



**SYMPOSIUM ON SPEECH COMMUNICATION  
METRICS AND HUMAN PERFORMANCE**

**Charles W. Nixon, PhD, Compiler**

**CREW SYSTEMS DIRECTORATE  
BIODYNAMICS AND BIOCOMMUNICATIONS DIVISION  
WRIGHT-PATTERSON AIR FORCE BASE OHIO 45433-7901**

**19961022 079**

**MAY 1995**

**DTIC QUALITY INSPECTED 1**

**INTERIM REPORT FOR THE PERIOD MAY 1993 TO MAY 1995**

**Approved for public release; distribution is unlimited**

**AIR FORCE MATERIEL COMMAND  
WRIGHT-PATTERSON AIR FORCE BASE, OHIO 45433-6573**

**ARMSTRONG  
LABORATORY**

## NOTICES

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise, as in any manner, licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use or sell any patented invention that may in any way be related thereto.

Please do not request copies of this report from the Armstrong Laboratory. Additional copies may be purchased from:

National Technical Information Service  
5285 Port Royal Road  
Springfield VA 22161

Federal Government agencies and their contractors registered with Defense Technical Information Center should direct requests for copies of this report to:

Defense Technical Information Center  
8725 John J. Kingman Rd., STE 0944  
Ft Belvoir VA 22060-6218

### TECHNICAL REVIEW AND APPROVAL

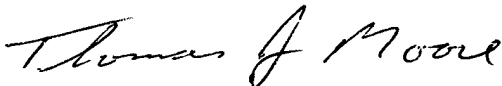
AL/CF-SR-1995-0023

This Special Report is published as received and has not been edited by the technical editing staff of the Armstrong Laboratory.

This report has been reviewed by the Office of Public Affairs (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public, including foreign nations.

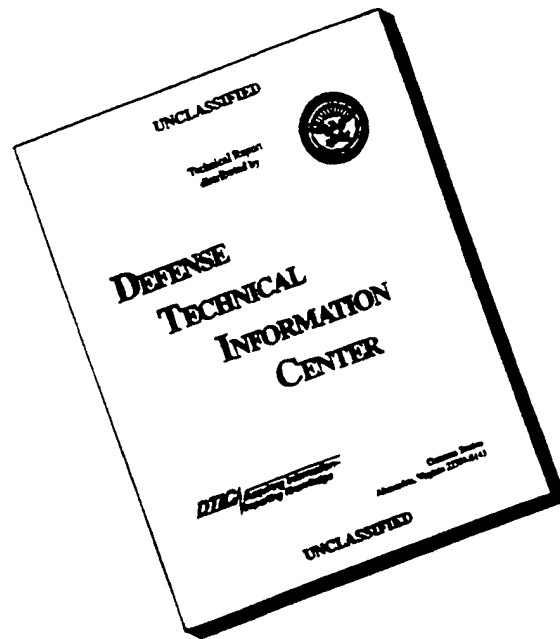
This technical report has been reviewed and is approved for publication.

FOR THE DIRECTOR



THOMAS J. MOORE, Chief  
Biodynamics and Biocommunications Division  
Crew Systems Directorate  
Armstrong Laboratory

# DISCLAIMER NOTICE



**THIS DOCUMENT IS BEST  
QUALITY AVAILABLE. THE  
COPY FURNISHED TO DTIC  
CONTAINED A SIGNIFICANT  
NUMBER OF PAGES WHICH DO  
NOT REPRODUCE LEGIBLY.**

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.				
1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE May 1995		3. REPORT TYPE AND DATES COVERED Interim - 1 May 1993 to 1 May 1995
4. TITLE AND SUBTITLE Symposium on Speech Communication Metrics and Human Performance			5. FUNDING NUMBERS PE - 62202F PR - 7231 TA - 21 WU - 04	
6. AUTHOR(S) Charles W. Nixon, PhD, Compiler				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Armstrong Laboratory, Crew Systems Directorate Biodynamics and Biocommunications Division Human Systems Center Air Force Materiel Command Wright-Patterson AFB OH 45433-7901			8. PERFORMING ORGANIZATION REPORT NUMBER  AL/CF-SR-1995-0023	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)			10. SPONSORING/MONITORING AGENCY REPORT NUMBER	
11. SUPPLEMENTARY NOTES				
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited			12b. DISTRIBUTION CODE	
13. ABSTRACT (Maximum 200 words) Various metrics are used to evaluate different aspects of speech communication. Most popular are measures of speech intelligibility. Other metrics that measure more than intelligibility are being employed, but they are not well known or readily available to potential users. This report contains reviews of some current practices and new research on metrics used to quantify speech communication. Three areas of speech communication are reported. The first reviews the development of speech intelligibility measures, the American National Standards Institute, S3.2-1989, "Method for the Measurement of Intelligibility of Speech over Communication Systems," and the proposed revision of the Articulation Index. The second addresses relationships between speech intelligibility and task performance showing that the complexity of both the speech and of the task affects performance. Some procedures allow determination of the amount of intelligibility required to complete specific tasks. The third area addresses the roles of multiple modalities in speech perception, primarily the auditory and visual, in areas of integration models, generalizability theory, tests of basic discrimination abilities, and sequence comparison measurement techniques. Additional information on these topics is available from the individual researchers.				
14. SUBJECT TERMS Speech intelligibility, speech metrics, task performance and speech, variability in speech tests, individual differences in speech reception			15. NUMBER OF PAGES 154	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT UNCLASSIFIED	18. SECURITY CLASSIFICATION OF THIS PAGE UNCLASSIFIED	19. SECURITY CLASSIFICATION OF ABSTRACT UNCLASSIFIED	20. LIMITATION OF ABSTRACT UNLIMITED	

This page intentionally left blank.

## PREFACE

This special report is a compilation of the presentations made at the Symposium on Speech Communication Metrics and Human Performance held 3-4 June 1993 at the National Academy of Sciences, Washington DC. The symposium was sponsored by the National Academy of Sciences - National Research Council (NAS-NRC) Committee on Hearing, Bioacoustics, and Biomechanics (CHABA). The program was organized by Thomas J. Moore (Chair), Judy R. Dubno, and Neal F. Viemeister of that committee. The organization and operation of the symposium were accomplished by the NAS-NRC staff associated with the CHABA, with special support provided by Arlyss K. Wiggins, Senior Program Assistant.

The topic of the symposium for the 1993 annual meeting was selected by the committee in response to a proposal from the Air Force. Symposium participants provided manuscripts of their presentations for inclusion in a final report. No reviews or editing of the manuscripts provided by the authors were accomplished in the preparation and publication of the report.

The symposium was one of the last ones sponsored by CHABA and one of the committee's last activities. NAS-NRC work in this scientific area is now being developed under a new Task Force on Behavioral, Cognitive, and Sensory Sciences.

The information contained in this report supports task 723121, Voice Communications, Work Unit 72312104, Bioacoustics and Biocommunications Research, Bioacoustics and Biocommunications Branch, Biodynamics and Biocommunications Division, Crew Systems Directorate, Armstrong Laboratory, Wright-Patterson AFB, OH. Technical support for the preparation of the report was provided by Merry Spahr and Marty Luka of Systems Research Laboratories, Inc.

This page intentionally left blank.

## TABLE OF CONTENTS

INTRODUCTION .....	1
DEVELOPMENT OF SPEECH INTELLIGIBILITY MEASURES AND THE ANSI STANDARD .....	3
PROPOSED REVISION OF ANSI STANDARD S3.5 FOR CALCULATION OF THE ARTICULATION INDEX .....	14
SOURCES OF VARIABILITY AFFECTING SPEECH PERCEPTION AND SPOKEN WORD RECOGNITION .....	19
SPEECH INTELLIGIBILITY EFFECTS IN A DUAL-TASK ENVIRONMENT .....	41
THE EFFECTS OF SPEECH INTELLIGIBILITY ON MILITARY PERFORMANCE .....	72
DEVELOPING A MESSAGE COMPLEXITY INDEX .....	83
INDIVIDUAL DIFFERENCES IN SPEECH PERCEPTION BY EYE AND EAR .....	90
SEQUENCE COMPARISON TECHNIQUES CAN BE USED TO STUDY SPEECH PERCEPTION .....	103
APPLICATIONS OF GENERALIZABILITY THEORY TO MEASUREMENT OF INDIVIDUAL DIFFERENCES IN SPEECH PERCEPTION .....	123
INTEGRATION MODELS OF SPEECH INTELLIGIBILITY .....	129
APPENDIX A. PROGRAM OF THE SYMPOSIUM ON SPEECH COMMUNICATION METRICS AND HUMAN PERFORMANCE .....	145



This page intentionally left blank.

## INTRODUCTION

Various metrics are used to evaluate different aspects of speech communication. Most popular are those that measure speech intelligibility or recognition. Other metrics are being employed that measure more than intelligibility; however, they are not well known or readily available to potential users. This situation, as well as the recent Committee on Hearing, Bioacoustics, and Biomechanics (CHABA) meeting on speech enhancement, indicated the need for a review of some current practices and an update of relatively new techniques and metrics used to quantify speech communication.

The CHABA annual scientific meeting in 1993 provided the forum in which the review and the update of "new" metrics were accomplished. Active researchers explained the assumptions, methodologies, and applications of their metrics during the symposium (symposium program is shown in Appendix A). Manuscripts addressing the research which has been conducted on these metrics were provided for publication. This special report is comprised of those manuscripts.

The initial session of the symposium reviewed the development of speech intelligibility measures and of the American National Standards Institute (ANSI) standard S3.2-1989, "The Measurement of the Intelligibility of Speech over Communication Systems." A proposed revision of the ANSI standard for determining the Articulation Index (AI) was presented. The AI is a physically-based metric that is predictive of the intelligibility of speech. Sources of variability affecting speech intelligibility tests were explained with the suggestion that long-standing theoretical issues in speech perception be reassessed and new directions for models of speech perception be pursued.

The effects of different levels of speech intelligibility on task performance were addressed in the second session of the symposium. Although it is self-evident that speech intelligibility has an effect on task performance, these new metrics enable these effects to be quantified. The complexity of both the messages and the tasks affects performance. As message complexity increases, both task errors and task response times increase. As task complexity increases, performance begins to decrease at high speech intelligibility levels. An information-theory-based metric includes such variables as vocabulary, task, time windows, and acoustic environments of communicators in the determination of the relative effectiveness of the communications. It also allows determination of the amount of speech intelligibility required to complete a specific task.

The roles of multiple modalities in speech perception, primarily auditory and visual, were addressed in the third session. Cross-modal "optimum processor" integration models of speech intelligibility show that the perception of speech is determined by the integration of cues, such as acoustic and visual, from several sources. Evidence is also presented of ear and eye modality-independent sources of variance in speech perception. The application of the generalizability theory in the measurement of individual differences permits estimations of the effects of several extraneous variables and their interactions in a single experiment. Sequence comparison

techniques provide numerical measures of phonemes correct; insertions, deletions, and substitutions; and can be applied to auditory and audio-visual connected speech.

The body of this publication contains the manuscripts provided by the researchers in the order in which they were presented at the symposium. The individual participants should be contacted if additional information is desired on any of the topics.

# **DEVELOPMENT OF SPEECH INTELLIGIBILITY MEASURES AND THE ANSI STANDARD**

**Mones E. Hawley**  
**Jack Faucett Associates**

## **INTRODUCTION**

### **SPEECH INTELLIGIBILITY MEASUREMENTS BEFORE 1940**

Investigations before 1900  
Investigations between 1900 and 1940

### **SPEECH INTELLIGIBILITY MEASUREMENTS SINCE 1940**

Investigation in Telephony  
Investigations in Architecture  
Investigations in Audiometry

## **STANDARDIZATION OF INTELLIGIBILITY TESTING**

---

## **INTRODUCTION**

I intend to start with a deadly serious story.

In 1942 I enlisted in the Army and became a Norden bombsight mechanic. I was assigned to a unit with airplanes that did not use such sophisticated bombsights and began flying night practice missions as a bombardier while I studied aerial navigation. In Egypt, Libya, and Tunisia, I was a navigator on courier and compass calibration flights. Just about 50 years ago today, I began to fly combat missions as a bombardier, then as a navigator-bombardier, and finally as the lead navigator when our squadron led the group of 48 aircraft.

Our aircraft were twin-engined B-25s, probably the noisiest combat airplanes used by the Army Air Force. The navigator's position was only a few inches behind the plane of the propellers, the noisiest place in the airplane. We had carbon microphones and used throat microphones when both hands needed to be busy. Both the ambient noise levels and the distortion in the communication system were very high.

In February 1944, on my 35th combat mission, I was shot down over the Anzio beachhead because the other crew members could not understand my directions over the interphone. I was wounded, parachuted out, and was taken prisoner. The rest of my crew was killed.

Six years later at RCA, I found myself, only partially accidentally, inventing and developing microphones for high intelligibility communications on naval ships and in aircraft. Over a few years, I expanded my interests in the intelligibility of whole communications systems to include the talkers and listeners, their environments, and their equipment, both electroacoustic and electronic. The evaluation and the prediction of that intelligibility are the subject of the first two papers.

It is only a few days since Memorial Day, and I should like this talk to be a particularly good one to honor my crew of 50 years ago.

## **SPEECH INTELLIGIBILITY MEASUREMENTS BEFORE 1940**

### **Investigations before 1900**

Concern for the intelligibility of speech is as old as speech itself. Speech is a fine example of adaptive feedback among humans. The ear does not perceive as speech what the speech mechanism cannot produce reliably, and the speech mechanism does not intentionally produce sounds which the ear cannot perceive or distinguish.

The ancients designed amphitheaters and, later, architects designed churches and theaters so the audiences could see and sometimes hear the words of the actors and singers. Pulpits and daises added both visibility and audibility. Sometimes resonators and reflectors were added. Ear trumpets and megaphones were developed to increase the sound pressure level at the ears of listeners. Whisper and spoken word tests were used to measure hearing acuity as early as the 16th century.

However, it was the invention of the telephone that began the science of intelligibility and of intelligibility testing as we know it today. In 1876 Alexander Graham Bell wrote:

"Indeed as a general rule, the articulation was unintelligible except when familiar sentences were employed....The elementary sounds of the English language were uttered successively into one of the telephones and its effects noted at the other. Consonantal sounds, with the exception of L and M, were unrecognizable. Vowels sounds in most cases were distinct."

In this paper Bell established a precedent for measuring the intelligibility of articulation by means of isolated speech sounds. I shall use the term intelligibility to mean the ability of listeners without hearing impairments to recognize speech units spoken by talkers without articulation impairments. I shall try to use the term discrimination to refer to hearing tests using words.

Edison invented the phonograph only a few years after Bell invented the telephone. In 1889 Lichtwitz, a German physician practicing in Prague, published papers in French and in German on a speech audiometer using a phonograph. He did his testing outdoors and varied the level of the stimulus by removing the subject to greater and greater distances. He wrote:

"With the assistance of this apparatus [the Edison phonograph] it will be possible to prepare some phonograms which can serve as acoumetric scales according to the model of optometric scales. These scales will bear the impressions of vowels, consonants, syllables, words, and phrases according to their intensities...with the scales one will be able to examine the ear from the point of view of perception of the selected sounds and noise....It is thus possible to let our patients hear the same phonogram reproduced every time in the same way."

Lichtwitz's paper seems to have been forgotten. Later developers of speech audiometers do not refer to it, and, for the next half century, the major research in intelligibility was conducted by engineers and scientists concerned with telephony.

### **Investigations between 1900 and 1940**

During the first 40 years of this century, the intelligibility investigators were usually trained as physicists. Electronic engineering had not become a major field of study, and psychologists did not often turn to audition as a specialty. At the beginning of the century a speech laboratory looked like Professor Higgins' facility in the film Pygmalion. The vacuum tube, especially when used as an amplifier, changed all that. The vacuum tube amplifier was the major tool that led to complex communications equipment, recording equipment, and instruments that permitted complicated experiments that could be replicated. Speech could be presented at different levels without changing the speaking level or spectrum. Spectra could be shaped and cut off. Shaped noise spectra could be introduced in controlled amounts. Signal-to-noise ratios and speech peaks could be measured reliably. The introduction of the oscilloscope toward the end of this period permitted speech scientists to visualize the instantaneous sound pressure changes that previously had been described only in equations. Electronics became the language of speech experimenters.

The telephone laboratories of the world led in most of the speech intelligibility investigations. Of these the Bell Telephone Laboratories was unquestionably the most important. In 1910 George A. Campbell published a paper in which he reported Bell Labs' intense concern with the recognition of the individual sounds of speech. It was assumed that there were 30-50 of these and that they could be strung together like a line of type. He quotes from a paper by Lord Rayleigh in which he describes confusion between the f sound and the s sound, and says that when a talker counted "one, two, three, four, sive, fix" over the telephone, the sequence of numbers sounded just as it did when the words were pronounced correctly.

Speech scientists today should note that, although European telephone laboratories were part of their national Post-Telegraph-Telephone system, there was almost no Federal involvement in speech research in this country until World War II. The research was funded by private companies and by universities. RCA and Western Electric were the dominant makers of high quality microphones, amplifiers, recording equipment, and sound-on-film equipment for the movies. NBC, CBS, and Western Electric developed the standard for the VU meter without anyone else's help or approval.

First Irving Crandall and then Harvey Fletcher led an outstanding group of scientists at Bell Labs who discovered many of the phenomena and rules of behavior on which speech research was based, for example:

- 1925 C. F. Sacia defined and measured speech power factors
- 1929 Fletcher and Steinberg showed sentence vs. syllable intelligibility
- 1934 John Steinberg clearly foreshadowed the sound spectrogram
- 1938 A. H. Inglis introduced the concept of orthotelephonic gain
- 1939 Homer Dudley developed the first important speech synthesizer
- 1940 Dunn and White established the basis for the AI

By 1940 the effects on intelligibility of filtering and of signal-to-noise ratio separately and taken together were pretty well understood. The basis had been established for the investigators who contributed so much during and immediately after World War II.

Before turning to that period, we should note the contributions of Vern Knudsen at UCLA to speech intelligibility in buildings. In 1929 he published a paper in volume 1 of the Journal of the Acoustical Society of America entitled "The Hearing of Speech in Auditoriums" in which he describes intelligibility testing. His classic book, Architectural Acoustics, published in 1932, describes the application of nonsense syllable testing and calculation of an index from the vowel and consonant scores. Vermeulen in Holland also did experiments on speech intelligibility in auditoriums.

Speech audiometry proceeded slowly during this 40-year period. Bryant invented a speech audiometer in 1904 and modestly claimed:

"The phonograph acoumeter invented by me overcomes all difficulties, for it can be manufactured in large numbers with perfect accuracy, and the pitch and intensity of its mechanical human voice do not vary....Records are made from carefully selected monosyllabic words in common use, with special reference to the logographic value of their consonants."

Bryant's attenuators were mechanical valves in tubes leading from the phonograph to the patient's ears. The device was not a success. The first widely-used speech audiometer was the Western Electric 4A introduced in the 1930s. In the United Kingdom, several investigators devised carrier sentences and lists of words and syllables for testing different populations. In this country, Alfred Wengel developed a CVC rhyme test for testing hearing. However, in audiometry there was nothing like the concentration of research there was in telephony.

## **SPEECH INTELLIGIBILITY MEASUREMENTS SINCE 1940**

### **Investigations in Telephony**

During World War II, intelligible speech communications meant the difference between life and death. All the combatants had noisy engine rooms, armored vehicles, and aircraft, but only the United States spent sizeable scientific resources on solving the problems. The problems

had been experienced in World War I. Two patents for noise-canceling microphones, issued in 1917 and 1919, read, in part:

"The need for a telephone system which will clearly and distinctly transmit speech from an observer to a pilot, or vice versa, in an aeroplane is especially crying and since the beginning of the present war inventors and research workers everywhere have been striving to produce such an apparatus."

"The invention was suggested by the necessity of telephoning in or from airplanes where the noise of the engines interferes seriously with the operation of ordinary telephone transmitters."

Several laboratories were established at universities and at military training stations especially to address the problems. The two most famous were at Harvard University: the Electro-Acoustic Laboratory directed by Leo Beranek and the Psycho-Acoustic Laboratory directed by S. S. Stevens. I believe that there are two reasons why psychologists became major researchers in this field during this period. The first reason is that experimental psychologists were becoming used to using electronic instruments in experiments. The second is that the physicists and engineers were much in demand for the radar, sonar, communications, and nuclear weapons programs. The wartime work on intelligibility was assigned to the acoustics section of the physics division of the Office of Scientific Research and Development. The division head was Harvey Fletcher. His text, Speech and Hearing, published in 1929, is a classic in the field.

By the end of World War II, there were three major results of the wartime work on intelligibility. First, America was clearly the world leader in this field of science. Second, there was an enormous reserve of publishable research and development on understanding and improving speech intelligibility over communications equipment. Finally, there were many trained scientists and engineers who, quite naturally, wanted to continue their investigations. In the period 1945-1955, there was an outpouring of research papers that has not been equaled since. Almost all the effects that interfere with speech intelligibility were investigated quite thoroughly. Bandwidth limitations, noise, distortion, peak and center clipping, interruption, and reverberation were studied, and papers were published that must be read by investigators today. Half the notable papers on intelligibility published in that period had as senior author a scientist who worked at the Harvard Psycho-Acoustic Laboratory.

The tape recorder was the most important item of new equipment. It permitted precise replication of stimuli and modification of arrangement or sequence or duration of the stimuli. It also permitted recording of results or environments in the field for later analysis and experimentation in the laboratory. Tape loops and re-recording techniques permitted the experimenter to create almost any combination of speech and interference or distortion. The stored-program digital computer began as a calculator, then became a controller of experiments and an analyzer of speech intelligibility results. Eventually, of course, it became a substitute for subjects and processors.



In the field of telephony, the United States continued to be the world leader in research. Federally-sponsored research and development programs supported new military and space vehicles. Now, however, the work was widely diffused throughout military, commercial, and academic laboratories, and none dominated the field. European nations rebuilt their laboratories and became familiar with the American wartime research. Then they began to contribute original research, particularly in the health and architectural applications.

### **Investigations in Architecture**

Because reverberation was not a common problem in telephony, this form of distortion was not investigated very thoroughly by communication equipment scientists. Bolt and MacDonald published a seminal paper on this subject in 1949. European and Japanese scientists made major contributions to this field and to the measurement of room characteristics that affect speech. Noise control engineers investigated the design of offices and other work places to determine the effects of fans, ventilation ducts, and suspended ceilings on speech intelligibility. The open, landscaped office and ubiquitous fan-cooled electronic office equipment presented new problems for speech privacy and intelligibility. Intelligibility testing of buildings and auditoriums was largely avoided because of the expense of using a number of subjects large enough to draw significant conclusions.

### **Investigations in Audiometry**

During the post-war period, there was a great deal of activity in this country and abroad in the investigation of the intelligibility of speech by people with hearing impairments. Part of this activity was the result of the concern for injured war veterans and the increasing noisiness of military vehicles, but most of the research was directed at the general population. The research was applied to several fields:

- \* factors affecting intelligibility for listeners with impaired hearing
- \* prediction of speech reception thresholds
- \* discrimination losses from pure tone audiograms
- \* testing materials in various national and regional languages
- \* diagnostic testing materials and techniques
- \* tests for malingering
- \* measurements to help select hearing aids
- \* comparison of monaural and binaural hearing aids
- \* standardization and validation of tests and equipment

In my view, some of the most important progress was made in the increase in rigor of the experiments and the testing procedures. Hallowell Davis and Ira Hirsh at the Central Institute for the Deaf deserve special mention for their contributions to this field. In 1948 Davis suggested a Social Adequacy Index based on a listener's ability to hear test words presented at a normal speaking level. Lichtwitz, in the paper I cited earlier, quotes D. B. S. Roosa in 1885:

"After the child had submitted with good grace to the lengthy examination of his hearing by means of watch ticks, he finally asked, 'What difference does it make whether I hear the watch, I want to hear what people say to me.'"

In 1952 Hirsh and his colleagues published "Development of Materials for Speech Audiometry" and made their recorded W-2 word lists widely available. Today, with tens of hundreds of millions of dollars paid out annually in compensation for hearing loss damages, the evaluation of speech intelligibility is a major concern.

## STANDARDIZATION OF INTELLIGIBILITY TESTS

In this country, there have been three standardization efforts for communication systems:

- 1952 ASA Exploratory Group--produced a positive recommendation
- 1953 ASA Writing Group--produced ASA 3.2-1960, "American Standard Method for Measurement of Monosyllabic Word Intelligibility"
- 1979 ANSI Writing Group--produced ANSI 3.2-1989, "American National Standard Method for Measuring the Intelligibility of Speech Over Communication Systems"

The exploratory committee was chaired by Daniel Martin, and I was a member. Our assignment was to answer the questions, "Is it feasible to write a standard for testing speech intelligibility?" and "Is there a need for such a standard?"

The answer to both questions was affirmative, and the first writing committee was appointed with me as chairman. Seven years later we produced S3.2-1960. After the standard was almost 20 years old and had been reaffirmed several times, the Acoustical Society of America, acting for the American National Standards Institute, the new name, established another committee to bring the old standard up to date. I became chairman and the only person to serve on both committees. It took nine years, but we received only one negative vote on the letter ballot. That vote caused changes which improved the standard, and the vote was speedily reversed. In December 1989 the new S3.2 became an ANSI standard.

Both standards addressed the same users and the same purposes. The standard is intended for people who want quantified results of comparisons among alternative systems. The user I had in mind as I wrote part of this standard was an engineer preparing specifications for the communication system for a state police force. This engineer is well educated but not a specialist in audio engineering or psychoacoustics. He or she wants to ensure that the new system is at least as intelligible as the present system and that the measurements are made by an accepted method.

The 1960 standard used the Harvard phonetically-balanced (PB) word lists, 20 lists of 50 words each. The 1989 standard permits the use of the PB words or the modified rhyme test or the diagnostic rhyme test. Each standard confined itself to these materials because other alternatives had not been widely replicated and validated among several laboratories.

The 1989 standard specifies that at least five different talkers and at least five different listeners must be used. If the users are to be of both sexes, both sexes must be represented in the set of talkers and listeners. If children are to be users, they must be represented in the subjects. All the materials must be presented to all the listeners by all the talkers. All the stress conditions must be presented, too, but not all materials have to be presented under all stress conditions.

Our standard is not the only way to measure the intelligibility of speech. My favorite method was described by an early telephone researcher who said that a long time ago, before privacy laws, he and his colleagues listened in on overseas telephone calls and counted the number of times the listener asked the talker to repeat the message. The method is simple and has high face validity. Commercially available equipment uses another method, the RASTI, developed by Houtgast and Steeneken, to measure the intelligibility of speech in auditoriums where the primary stresses are reverberation and noise.

Figure 1 is reproduced from the ANSI standard. It shows a simple version of a speech communication system. Right now I am the talker, the sound system is the transmission medium, and you are the listeners. Some things that look simple are not. Talkers who distinguish "close" from "clothes" and "which" from "witch" are more intelligible than those who do not. If the talker is in very loud noise, the cavity of the mouth will have resonances which modulate the spectrum of the noise and change the signal-to-noise ratio. The effect is emphasized if the talker uses a noise-canceling microphone. This is one of the reasons speech intelligibility test results may be different if the talker is in noise rather than has noise added electrically to the speech signal.

If the talker is speaking in an enclosure, such as an oxygen mask or a space helmet, or if the talker is breathing a helium-oxygen mixture, the talker's environment may affect the articulatory movement or the speech spectrum. There are similar possible interactions between the listener and the listener's environment. There may be important differences between the environments for both the talker and the listener when the tests are conducted in the field or in a laboratory. For all these reasons, the new standard insists that the tests must be run in as realistic an operational environment as possible.

Figure 2 shows the middle of the previous diagram greatly expanded. This figure is in the new standard to show the additional stresses that may enter as one goes from face-to-face communication to analog intercoms to radio links, and then to coded, digital, and encrypted speech signals. The diagram is included in the standard to remind the experimenter to report all the conditions of the communication system and the stresses.

There are several groups for which the ANSI standard is not intended. One group is speech researchers. They already know how to design their experiments, to select and to train their subjects, and to report their results to be useful to other researchers. They may use the standard for a reference or a check list, but we did not write it for them.

A second exclusion concerns systems which are used by people who are not trained to do so. A person who listens to a public address system in a building with which he or she is not

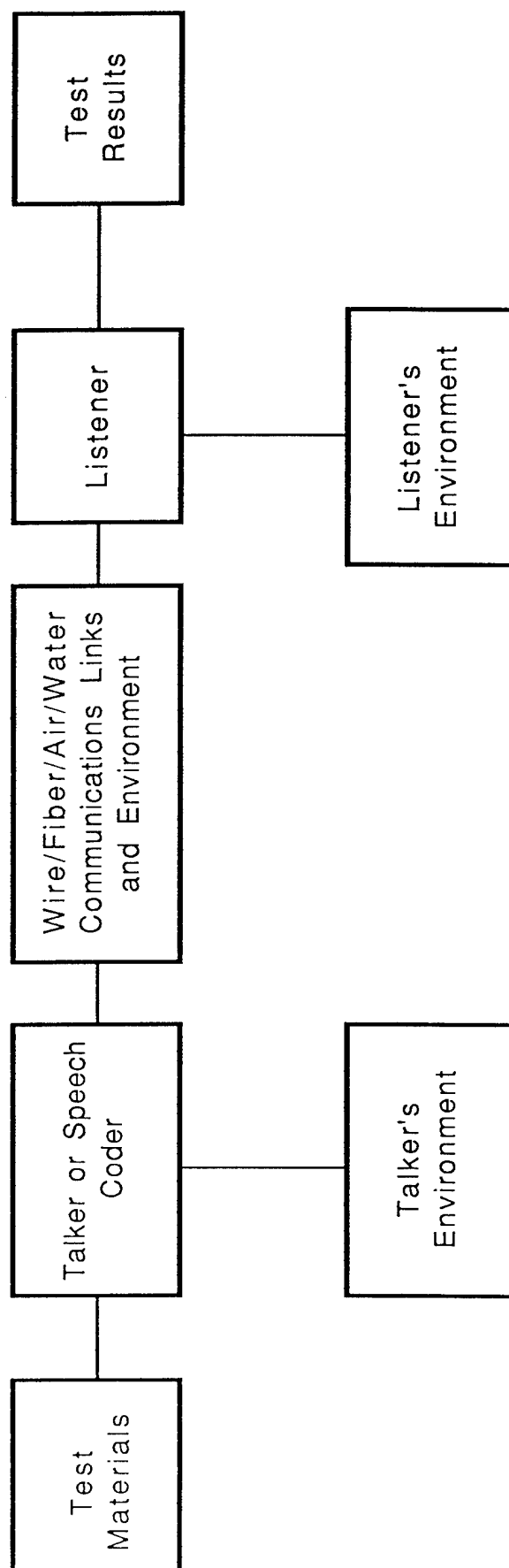


Figure 1

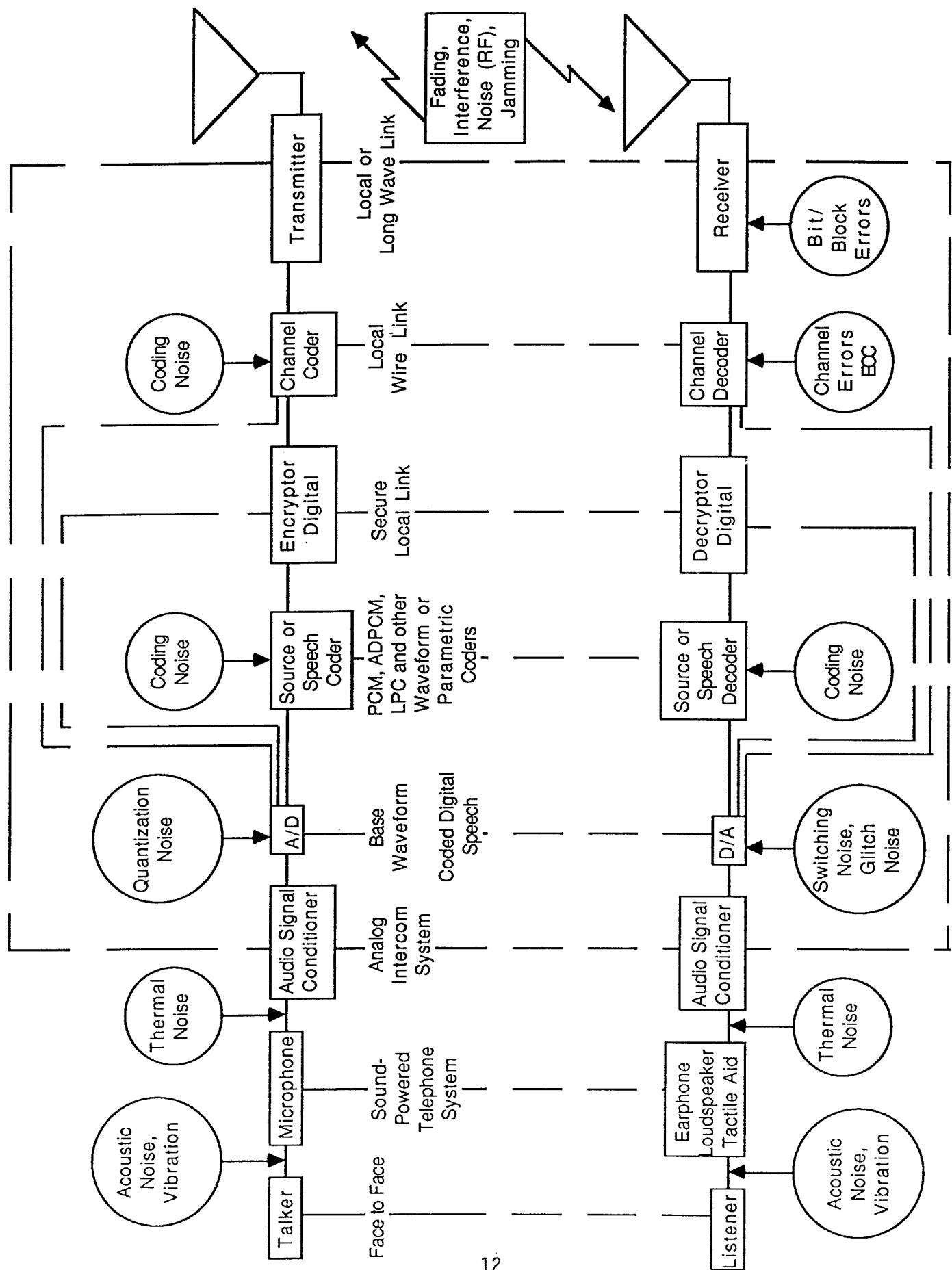


Figure 2

familiar is in this situation. A person who listens for the first time to synthesized speech over the telephone is in this situation. The standard we prepared is not suitable for realistic operational evaluation of such systems. A subject is not truly naive after the first word is heard. One immediately is measuring the state of learning instead of differences in systems. Listen to this short recording of the ground and local controllers at Dallas-Fort Worth International talking to aircraft. The FAA recorded this last year and it is representative of the state of speech communications today. Our standard is intended to evaluate such systems only when the listeners have become fully acquainted with the characteristics of the system and of the talkers.

The standard is intended for evaluating the intelligibility of speech over communication systems when used by human talkers and listeners without impairments. None of us on the committee was expert at preparing tests for articulating or discriminating speech.

Also, the standards are not intended to be used for testing speech-by-rule synthesizers or speech analyzers. Our committee was not ready to attempt standardization of testing of these devices.

Finally, the ANSI standard is not for testing speaker identification or recognition devices or for measuring speech quality. The standard is not intended to be used to measure emotion or mental state. The standard is intended for users of American English, although we know of no reason why other versions of English or other languages will not find it valuable if they have suitable speech materials. We hope our standard will be useful for the purposes for which we wrote it.

Thank you.

## **PROPOSED REVISION OF ANSI STANDARD S3.5 FOR CALCULATION OF THE ARTICULATION INDEX**

**C. V. Pavlovic**  
**Resound Corporation**  
**Redwood City California**  
**and**  
**University of Provence**  
**France**

**P. M. Zurek**  
**Research Laboratory of Electronics**  
**Massachusetts Institute of Technology**  
**Cambridge, Massachusetts**  
**and**  
**Sensimetrics Corporation**  
**Cambridge, Massachusetts**

The articulation index (AI) is a physically-based metric that is predictive of the effects of transmission degradations on the intelligibility of speech. AI was developed over the period from roughly 1910 to 1950, principally by researchers at the Bell Telephone Laboratories, and has been validated for a variety of intelligibility-degrading factors such as additive noise, filtering, and low speech presentation level. The methods for calculating AI that were set down by French and Steinberg (1947) and Beranek (1947) were simplified by Kryter (1962) and standardized in ANSI S3.5-1969.

Since 1969 there have been significant developments in the field of speech intelligibility. These have come not so much from further work in telephony, which motivated the original work, but from studies of intelligibility-prediction methods in architectural acoustics and from research into hearing impairment and hearing aids. The possibility of improving the validity of AI and extending its range of applicability motivated the formation of a committee\* of the Acoustical Society of America [Working Group S3.79] to draft a revision of ANSI S3.5-1969. This work, begun in 1986, is now (Fall, 1993) in the form of a draft that is being submitted for comment to the interested community at large. The following are the main changes being proposed, more or less in order of importance.

---

\* Members of Working Group S3.79, in addition to the authors, are: T. Bell, R. Bilger, A. Boothroyd, D. Dirks, J. Dubno, G. Garinther, M. Hawley, T. Houtgast, L. Humes, C. Kamm, K. Kryter, H. Levitt, G. Popelka, C. Rankovic, and G. Studebaker.

## **INCREASING THE APPLICATION DOMAIN OF THE STANDARD**

The most important changes to the standard relate to the need to provide a general framework into which various methods for determining the input variables (i.e., equivalent speech and noise spectra and equivalent hearing threshold levels) can be incorporated. For some applications, these methods already exist (e.g., the modulation transfer function for determining the apparent speech-to-noise ratio in reverberation), while others may still be developed. In addition, the generality of the standard has been extended to include various measurement points, such as free-field for architectural acoustics or eardrum for telephony. The revised standard is organized into two parts. Part I describes the calculation of the index when the input variables are known. The application domain of this framework is quite general and extends to all listening conditions where adequate methods for specifying these input variables exist. Part II gives measurement and calculation details for specifying these input variables for a number of conditions encountered in practice, such as external noise masking, reverberant speech, monaural listening, and some conditions of binaural listening.

The procedures for determination of the equivalent speech and noise spectrum levels (Part II) vary with respect to where and how the intervening variables are measured. The most general of the procedures requires measurement of the modulation transfer function for intensity (MTFI) and the combined speech and noise spectrum level (CSNSL) at the eardrum (MTFI and CSNSL are further discussed in the next section). Therefore, both the hardware for the MTFI/CSNSL measurement and a human head/ear mannequin are necessary. A less general procedure which is also discussed in detail excludes many communication situations (e.g., telephone links) and requires only the measurement of the MTFI and the CSNSL in a sound field at the position of the listener. The simplest of the discussed procedures requires only the measurement of the noise spectrum level in the absence of speech and an estimation (or measurement) of the speech spectrum level in the absence of noise. However, the field of application of this procedure is even further reduced, because conditions where reverberation decreases speech intelligibility and conditions where noise and speech spectrum levels depend on one another are excluded.

## **INCORPORATION AND ADAPTATION OF STI**

ANSI S3.5-1969 has only a crude correction for degradations due to reverberation. The division of the standard into the two parts discussed above makes it possible to include any existing procedure, or for that matter a future procedure, for the determination of the equivalent signal or noise level. Part II of the standard explains the utilization of the ideas employed in the Speech Transmission Index, or STI (Steeneken and Houtgast, 1980). The STI procedures have been modified to correspond to the AI bands, and the number of modulation frequencies has been decreased. Most importantly, the STI measurements have been supplemented with the measurement of the combined speech and noise spectrum level (CSNSL), extending the STI to conditions in which some part of the speech spectrum is inaudible. Finally, alternative procedures for measurement of the modulation transfer function for intensity (MTFI) are given. The modulation transfer function for intensity is the crucial variable determined by the STI.

At this point, it should be re-emphasized that it is not intended that the procedures elaborated in Part II, such as STI, be the only solutions endorsed by the standard. The user is



encouraged to employ any other technique more appropriate for the particular problem--as long as the necessary input variables for Part I are determined.

## **CHANGE IN IMPORTANCE WEIGHTS**

AI is a spectrally-weighted sum of effective band signal-to-noise ratios. In the original development of articulation theory, the frequency-weighting function was derived from careful studies of the intelligibility of filtered consonant-vowel-consonant syllables, and was thought to reflect the underlying importance of each frequency region to speech intelligibility. A number of studies from the 1950s onward, however, have shown that the early importance function may not be as fundamental as believed, at least not in as simple a way. The recent studies by Studebaker, Pavlovic, and Sherbecoe (1987) show most clearly that the importance function that is derived in an experiment depends on the test materials used. As the context inherent in the material increases (going from syllables to monosyllabic and polysyllabic words, and to sentences), the apparent importance shifts systematically to lower frequencies. Whereas a cutoff at about 1900 Hz divides the spectrum into equally-contributing halves for nonsense syllables, a similar cutoff for continuous discourse is found at about 1100 Hz.

Given that the interest in most applications of the articulation index is in high-context everyday speech, most closely approximated by sentences or continuous discourse tests, it was felt that the calculation of the revised index should reflect greater relative weighting for low frequencies than is currently given. The proposed revision adopts a compromise among various weighting functions, called the importance function for average speech, that was suggested by Pavlovic (1987). In addition, the Appendix to the standard gives the importance functions that have been derived for some common speech tests. These are useful in applications where the predicted AI is compared to the actual AI calculated from the measured speech intelligibility score.

## **CHANGE IN DYNAMIC RANGE OF SPEECH**

As was the case with the importance weights, recent studies have found that improved predictions result from use of a "dynamic range" different from that described in ANSI S3.5-1969. This dynamic range describes the limits on band signal-to-noise ratios for contribution to the index. Below a lower limit there is no contribution, and above an upper limit there is no change in contribution. In the current standard, these limits are -12 and +18 dB. Studies by Steeneken and Houtgast (1980) and Studebaker and Sherbecoe (1991) indicate that dynamic range limits of approximately -15 and +15 dB provide a better fit to data. These values have been adopted in the proposed revision.

## **CHANGE IN SPEECH SPECTRUM**

The speech spectrum has also been changed. The proposed speech spectrum has been obtained by averaging data from all available studies on the statistical distribution of speech. It is interesting to note that the proposed spectrum is roughly 3 dB below the ANSI S3.5-1969 spectrum, which, in the calculation of the AI, tends to cancel out the differences between the ANSI S3.5-1969 procedure and the proposed procedure due to the change in the dynamic range

of speech. Speech spectra are also given for a number of vocal efforts, ranging from normal conversational speech to shouted speech.

## **CHANGE IN CALCULATION BANDS**

ANSI S3.5-1969 provides three calculation procedures: an octave procedure, a 1/3-octave procedure, and an equally-contributing-band procedure (20 bands). The proposed standard retains the octave and the 1/3-octave procedures, but the 20 equally-contributing-band procedure has been replaced by an equally-contributing-critical-band procedure (17 bands). In addition, a somewhat more accurate critical-band procedure (not equally-contributing bands) is given.

## **SPEECH-LEVEL DISTORTION FACTOR**

The intelligibility of speech decreases gradually with increases in the presentation level of speech above a certain optimal level (French and Steinberg, 1947). The proposed standard introduces a speech-level distortion factor to account for this effect.

## **SELF-MASKING OF SPEECH**

For conditions of severe low-pass and high-pass filtering, ANSI S3.5-1969 could have produced an error because it does not account for the masking effects of lower speech frequencies on higher speech frequencies. The data of French and Steinberg were used to derive formulas that account for these masking effects.

## **CHANGE IN THE SPREAD OF MASKING PROCEDURES**

The spread of masking determination in ANSI S3.5-1969 is done graphically. The computational procedures described by Ludvigsen (1985) for 1/3-octave bands have been adapted for use in the revised standard.

## **AUTOMATED CALCULATION**

Many of the proposed changes discussed above (such as self-masking of speech and speech-level distortion factor) improve accuracy of the AI but they also render the calculations excessively complex for the manual chart-type computations used in ANSI S3.5-1969. The committee felt that the ubiquity of computers makes it now possible to depart from the manual chart-type computations and, instead, employ curve-fit equations which could be easily implemented in software. The basic AI calculations have been implemented in a program distributed with the proposed draft of the standard.

## **NAME CHANGE**

It was felt that this revision process provided a good opportunity to adopt a more descriptive name for the index. The index does not assess the quality of a talker's articulation. Rather, it summarizes the effects of degradations to the intelligibility of speech, without regard to how clearly it was produced. The proposed new name is the Speech Intelligibility Index.

In summary, it is hoped that these revisions will make the standard easier to use, more widely applicable, and more accurately predictive of everyday speech reception.

## REFERENCES

American National Standards Institute. "American National Standard methods for the calculation of the articulation index (ANSI S3.5-1969)" ANSI, New York (1969).

Beranek, L. "Design of speech communication systems," *IRE Proc.* 35, 880-890 (1947).

French, N.R., and Steinberg, J.C. "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, 19, 90-119 (1947).

Kryter, K.D. "Methods for the calculation and use of the articulation index," *J. Acoust. Soc. Am.*, 34, 1689-1697 (1962).

Ludvigsen, C. "Relations among some psychoacoustic parameters in normal and cochlearly impaired listeners" *J. Acoust. Soc. Am.*, 78, 1271-1280 (1985).

Pavlovic, C.V. "Derivation of primary parameters and procedures for use in speech intelligibility predictions" *J. Acoust. Soc. Am.*, 82, 413-422 (1987).

Steeneken, H.J.M., and Houtgast, T. "A physical method for measuring speech-transmission quality" *J. Acoust. Soc. Am.*, 67, 318-326 (1980).

Studebaker, G.A., Pavlovic, C.V., and Sherbecoe, R.L. "A frequency importance function for continuous discourse," *J. Acoust. Soc. Am.*, 81, 1130-1138 (1987).

Studebaker, G.A. and Sherbecoe, R.L. "Frequency-importance and transfer functions for recorded CID W-22 word lists," *J. Speech Hear. Res.*, 34, 427-438 (1991).

# **SOURCES OF VARIABILITY AFFECTING SPEECH PERCEPTION AND SPOKEN WORD RECOGNITION**

**David B. Pisoni**

**Speech Research Laboratory  
Indiana University  
Bloomington, Indiana 47405**

***Abstract.*** This paper reviews recent studies on the perception, encoding and retention of stimulus variability in speech perception. Experiments on talker variability, speaking rate and perceptual learning provide evidence for the encoding of very fine perceptual details of the speech signal. Listeners apparently encode specific attributes of the talker's voice and speaking rate into long-term memory. The process of perceptual normalization in speech perception therefore appears to involve the encoding of specific instances or episodes of the stimulus input and the processing operations used in perceptual analysis. The present set of findings is consistent with nonanalytic accounts of perception, memory and cognition which emphasize the contribution of episodic or exemplar-based encoding in long-term memory. The results also raise questions about the long-standing dissociation in phonetics between the linguistic and indexical properties of speech. Listeners apparently do encode and retain nonlinguistic information in long-term memory about the speaker's gender, dialect, speaking rate and emotional state, attributes of speech signals that are not traditionally considered part of phonetic or lexical properties of words. The findings reported here have important implications for current theoretical accounts of how the nervous system encodes speech signals and what kinds of information are stored in the mental lexicon.

## **SOURCES OF VARIABILITY AFFECTING SPEECH INTELLIGIBILITY TESTS**

For the last several years we have been interested in the interface between speech perception and spoken language comprehension and, in particular, problems of lexical access and the structure and organization of sound patterns in the mental lexicon (Pisoni, Nusbaum, Luce, and Slowiaczek, 1985). Findings from a variety of recent studies carried out at Indiana suggest that very fine details in the speech signal are preserved in the human memory system for relatively long periods of time (see Pisoni, 1990; 1992a,b, 1993; Goldinger, 1992). This information appears to be used in several ways to facilitate perceptual encoding, retention and retrieval of information from memory. Many of our recent investigations have been concerned with assessing the effects of different sources of variability in speech perception (Sommers, Nygaard, and Pisoni, 1992; Nygaard, Sommers, and Pisoni, 1992). The results of these studies have encouraged us to reassess our beliefs about several long-standing theoretical issues in speech perception such as acoustic-phonetic invariance and the problems of perceptual normalization (Pisoni, 1992).

In the sections below, I will summarize the results from several recent studies that deal with the encoding of stimulus variability in speech perception experiments. These findings have raised a number of important new questions about the traditional dissociation between the linguistic and indexical properties of speech signals and the role that different sources of variability play in speech perception and spoken word recognition. For many years, linguists have considered attributes of the talker's voice--what Ladefoged refers to as the personal characteristics of speech--to be independent of the linguistic content of the talker's message (Ladefoged, 1975; Laver and Trudgill, 1979). The dissociation of these two parallel sources of information in speech may have served a useful function in the formal linguistic analysis of language when viewed as an idealized abstract system of symbols. However, the artificial separation has at the same time created some difficult problems for researchers who wish to gain a detailed understanding of how the nervous system encodes speech signals and how real speakers and listeners deal with the enormous amount of acoustic variability in the speech signal.

## **EXPERIMENTS ON TALKER VARIABILITY IN SPEECH PERCEPTION**

Several novel experiments have been carried out to study the effects of different sources of variability on speech perception and spoken word recognition. We consider these studies to be novel because instead of reducing or eliminating variability in the stimulus materials, as most researchers have routinely done in the past, we specifically introduced variability from different talkers and different speaking rates to study their effects on perception (Pisoni, 1992). Our research on talker variability began with the observations of Mullennix et al. (1989) who found that the intelligibility of isolated spoken words presented in noise was affected by the number of talkers that were used to generate the test words in the stimulus ensemble. In one condition, all the words in a test list were produced by a single talker; in another condition, the words were produced by 15 different talkers, including both male and female voices. The results, which are shown in Figure 1, were very clear. Across three signal-to-noise ratios, identification performance was always better for words that were produced by a single talker than words produced by multiple talkers. Trial-to-trial variability in the speaker's voice apparently affects word recognition performance. This pattern was observed for both high-density (i.e., confusable) and low-density (i.e., non-confusable) words. Our findings in this study replicated results originally reported by Peters (1955) and Creelman (1957) back in the 1950s and suggested to us that the perceptual system must engage in some form of on-line recalibration each time a new voice is encountered during the set of test trials.

In a second experiment, we measured naming latencies to the same words presented in both test conditions (Mullennix et al., 1989). Table I provides a summary of the major results. We found that subjects were not only slower to name words from multiple-talker lists, but they were also less accurate when their performance was compared to naming words from single talker lists. Both sets of findings were again surprising to us at the time, because all the test words used in the experiment were highly intelligible when presented in the quiet. The intelligibility and naming data from these two experiments immediately raised a number of additional questions about how the various perceptual dimensions of the speech signal are processed by the human listener. At the time, we naturally assumed, as most people did in the past, that the acoustic attributes used to perceive voice quality were independent of the linguistic properties of the signal. However, to our knowledge no one had ever tested this assumption directly.

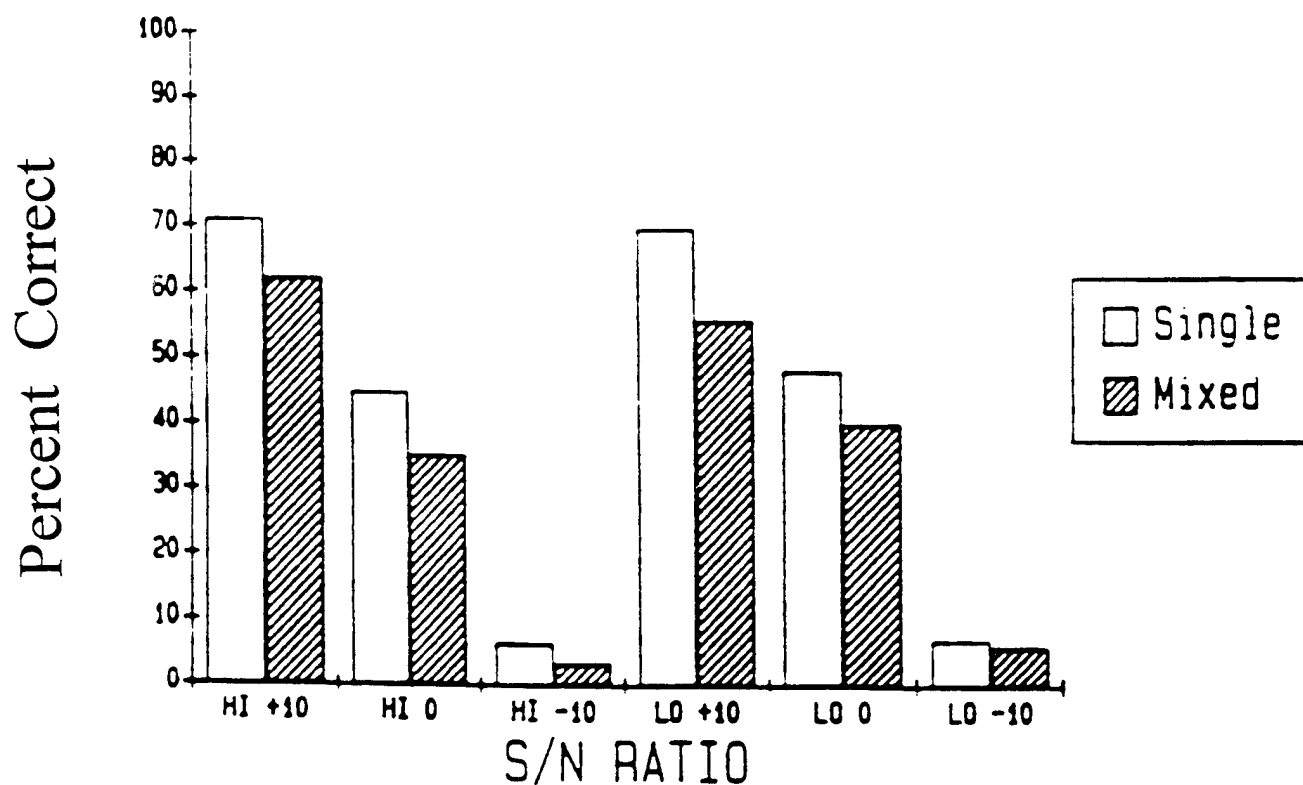


Figure 1. Overall mean percent correct performance collapsed over subjects for single- and mixed-talker conditions as a function of high- and low-density words and S/N ratio (from Mullennix et al., 1989).

Table I. Mean response latency (ms) for correct responses for single- and mixed-talker conditions as a function of lexical density (from Mullennix et al., 1989).

	Density	
	High	Low
Single talker	611.2	605.7
Mixed talker	677.2	679.4

In another series of experiments we used a speeded classification task (Garner, 1974) to assess whether attributes of a talker's voice were perceived independently of the phonetic form of the words (Mullennix and Pisoni, 1990). Subjects were required to attend selectively to one stimulus dimension (i.e., voice) while simultaneously ignoring another stimulus dimension (i.e., phoneme). Figure 2 shows the main findings. Across all conditions, we found increases in interference from both dimensions

when the subjects were required to attend selectively to only one of the stimulus dimensions. The pattern of results suggested that words and voices were processed as integral dimensions; the perception of one dimension (i.e., phoneme) affects classification of the other dimension (i.e., voice) and vice versa. Subjects apparently cannot selectively ignore irrelevant variation on the nonattended dimension. If both perceptual dimensions were processed separately and independently, as we had originally assumed, we should have found little if any interference from the nonattended dimension which could be selectively ignored without affecting performance on the attended dimension. Not only did we find mutual interference suggesting that the two sets of dimensions, voice and phoneme, are perceived in a mutually dependent manner, but we also found that the pattern of interference was asymmetrical. It was easier for subjects to ignore irrelevant variation in the phoneme dimension when their task was to classify the voice dimension than it was to ignore the voice dimension when they had to classify the phonemes in these stimuli.

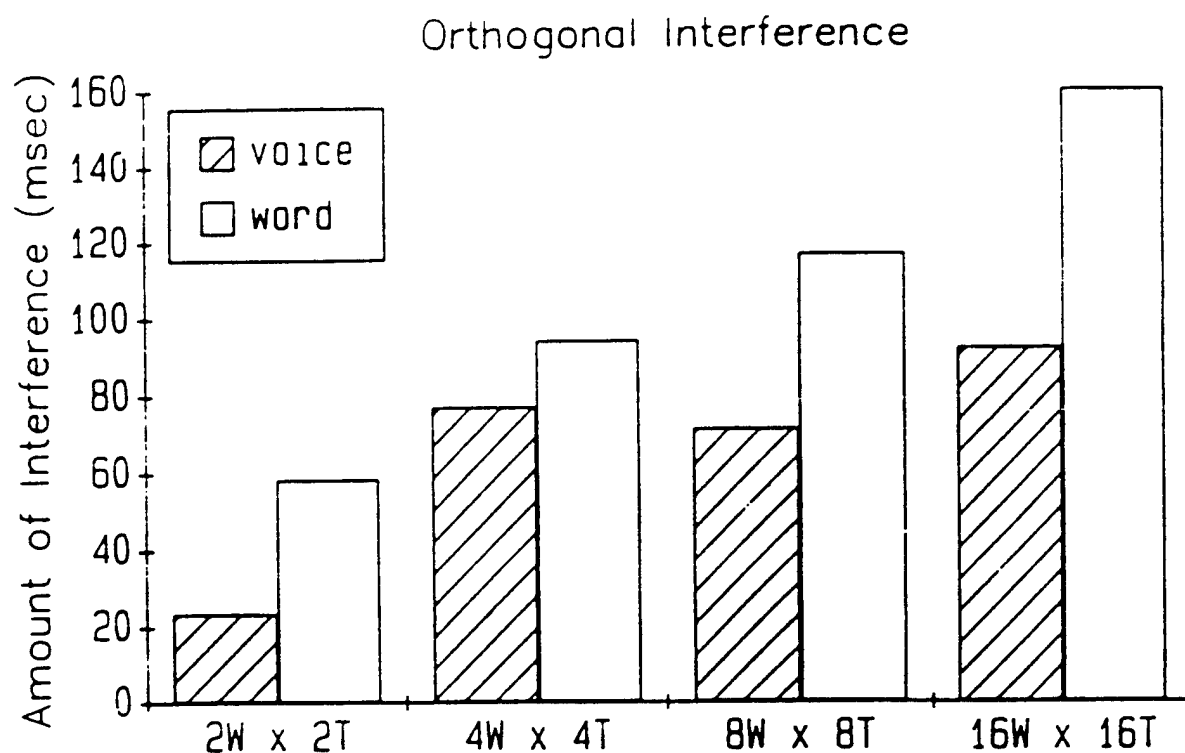


Figure 2. The amount of orthogonal interference (in milliseconds) across all stimulus variability conditions as a function of word and voice dimensions (from Mullennix and Pisoni, 1990).

The results from these perceptual experiments were surprising given our prior assumption that the indexical and linguistic properties of speech were perceived independently. To study this problem further, we carried out a series of memory experiments to assess the mental representation of spoken words in long-term memory. Experiments on serial recall of lists of spoken words by Martin et al. (1989) and Goldinger et al. (1991) demonstrated that specific details of a talker's voice are also encoded into long-term memory along with the to-be-remembered items. Using a continuous recognition memory procedure, Palmeri et al. (1993) found that detailed episodic information about a talker's voice is also encoded in memory and is available for explicit judgments even when a great deal

of competition from other voices is present in the test sequence. Palmeri et al.'s results are shown in Figure 3. The top panel shows the probability that an item was correctly recognized as a function of the number of talkers in the stimulus set; the bottom panel shows the probability of a correct recognition across different stimulus lags of intervening items. In both cases, the probability of correctly recognizing a word as old (filled circles) was greater if the word were repeated in the same voice than if it were repeated in a different voice of the same gender (open squares) or a different voice of a different gender (open triangles).

In another set of memory experiments, Goldinger (1992) found very strong evidence of implicit memory for attributes of a talker's voice which persists for a relatively long period of time after perceptual analysis has been completed. His results are shown in Figure 4. Goldinger also found that the degree of perceptual similarity affects the magnitude of the repetition effect in memory for identical voices, suggesting that the perceptual system encodes very detailed talker-specific information about spoken words in episodic memory representations.

Additional support for the proposal that detailed information about the talker's voice is encoded in memory comes from a recent experiment on sentence recall by Karl and Pisoni (1994). In this study, a cued recall procedure was used to study the retrieval of spoken sentences from long-term memory. After subjects transcribed lists of sentences, they were given a probed recall test with cues presented either visually or auditorily. Recall accuracy depended on the probe cues; when the probe words matched the study conditions, recall was highest, suggesting that detailed information about a talker's voice is encoded in long-term memory and that with the appropriate probe cue at the time of retrieval, this information can be accessed and used to recall the entire sentence.

Taken together, our recent findings on the effects of talker variability in perception and memory provide support for the proposal that detailed perceptual information about a talker's voice is preserved in long-term memory. At the present time, it is not clear whether there is one composite representation in memory or whether these different sets of attributes are encoded in parallel in separate representations (Eich, 1982; Hintzman, 1986). It is also not clear whether spoken words are encoded and represented in memory as a sequence of abstract symbolic phoneme-like units along with much more detailed episodic information about specific instances and the processing operations used in perceptual analysis. These are important questions for future research on spoken word recognition.

## **EXPERIMENTS ON THE EFFECTS OF SPEAKING RATE**

We also carried out another set of experiments to examine the effects of speaking rate on perception and memory. These studies, which were designed to parallel the earlier experiments on talker variability, have also shown that the perceptual details associated with differences in speaking rate are not lost or discarded as a result of perceptual analysis. In one experiment, Sommers et al. (1992) found that words produced at several different speaking rates (i.e., fast, medium and slow) were identified more poorly than the same words produced at only one speaking rate. These results were compared to another condition in which differences in amplitude were varied randomly from trial to trial in the test sequences. In this case, identification performance was not affected at all by variability in overall signal level. The results from both conditions are shown in Figures 5 and 6.



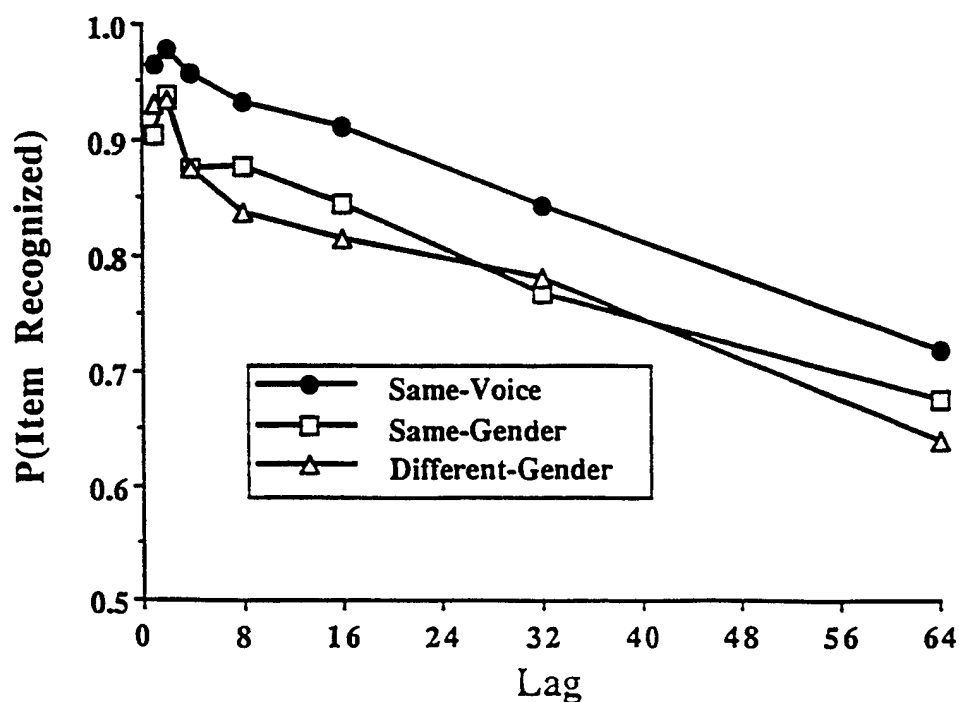
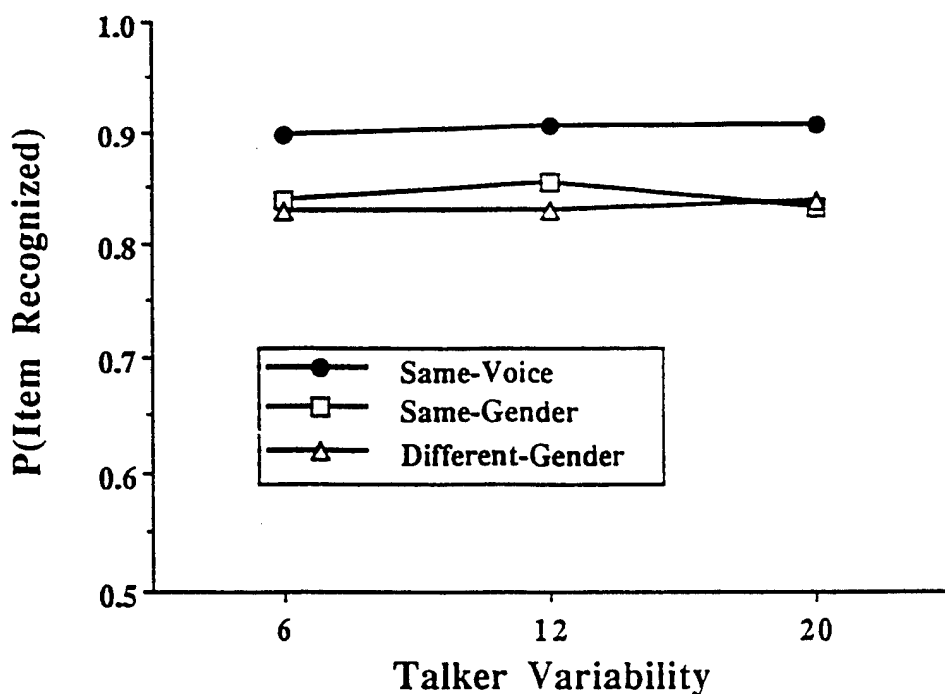
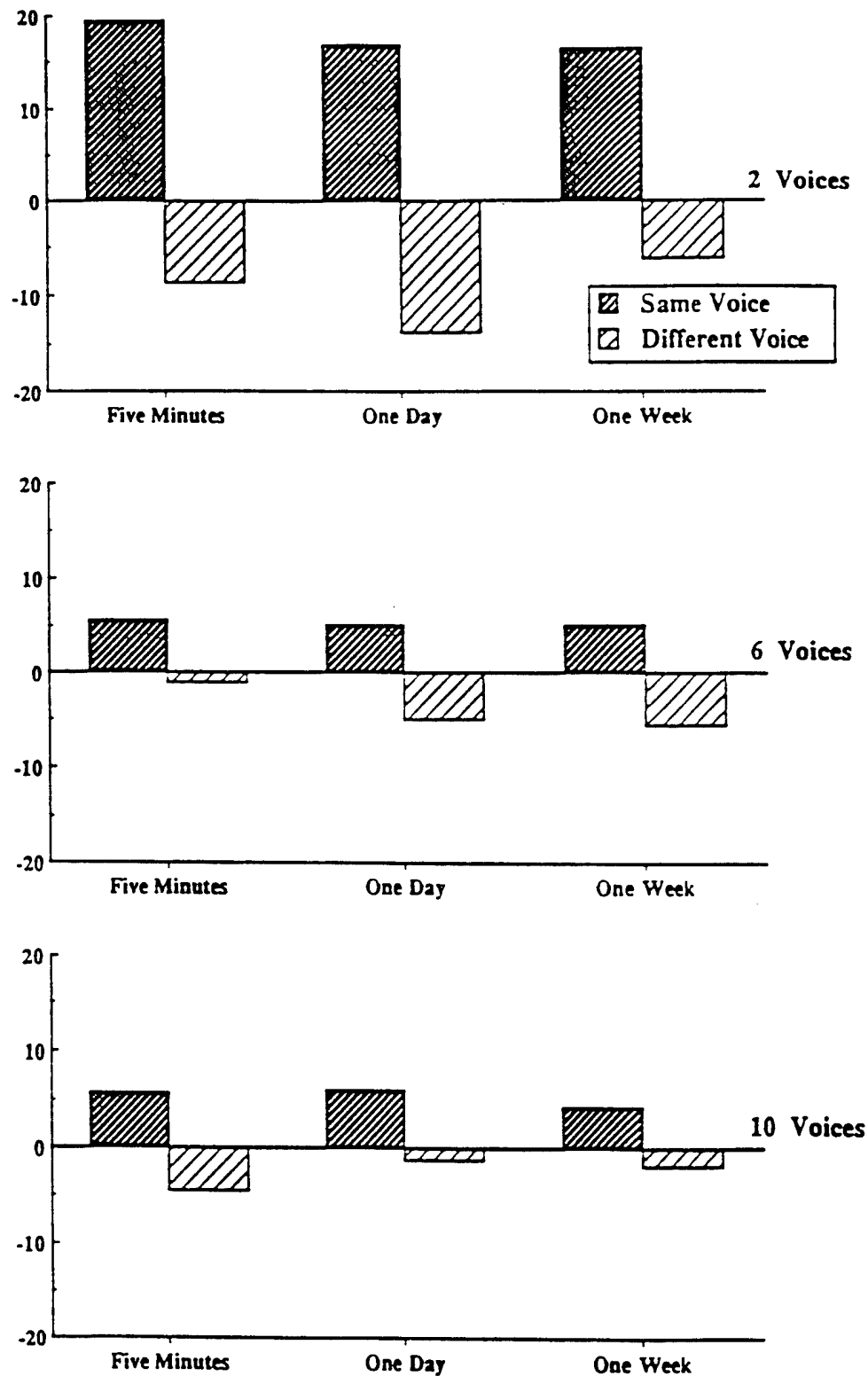


Figure 3. Probability of correctly recognizing old items in a continuous recognition memory experiment. In both panels, recognition for same-voice repetitions is compared to recognition for different-voice/same-gender and different-voice/different-gender repetitions. The upper panel displays item recognition as a function of talker variability, collapsed across values of lag; the lower panel displays item recognition as a function of lag, collapsed across levels of talker variability (from Palmeri et al., 1993).

# Net Repetition Effect (Percent)



## Delay Period

Figure 4. Net repetition effects observed in perceptual identification as a function of delay between sessions and repetition voice (from Goldinger, 1992).

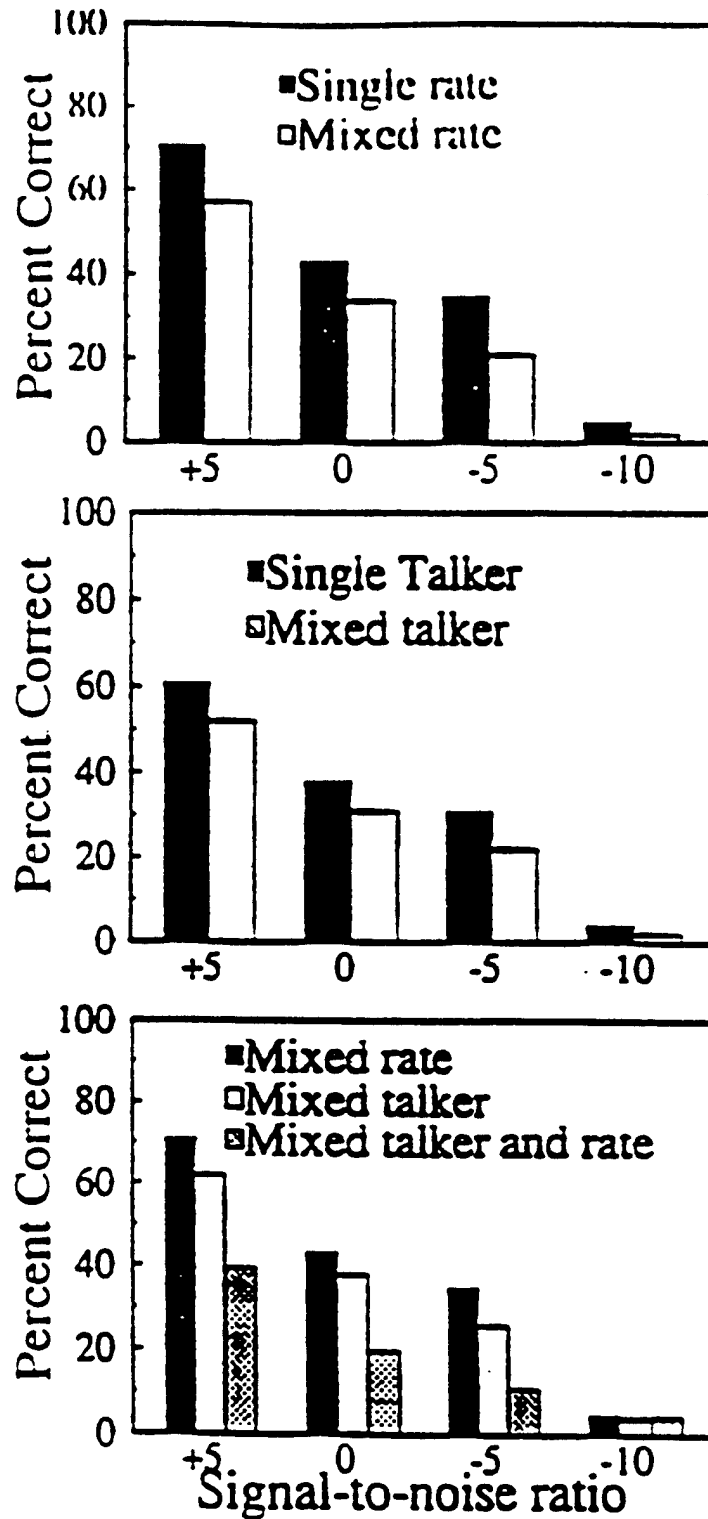


Figure 5. Effects of talker, rate, and combined talker and rate variability on perceptual identification (from Sommers et al., 1992).

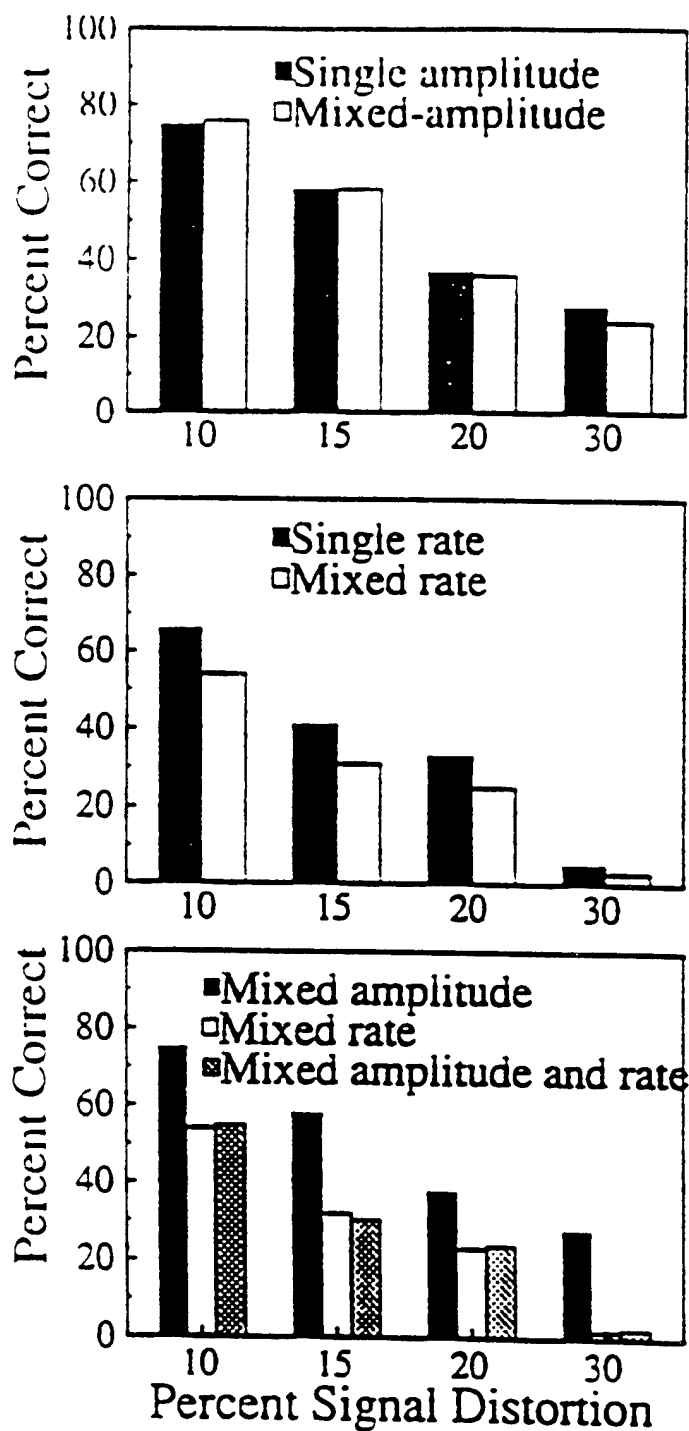


Figure 6. Effects of amplitude, rate, and combined amplitude and rate variability on perceptual identification (from Sommers et al., 1992).

Other experiments on serial recall have also been completed to examine the encoding and representation of speaking rate in memory. Nygaard et al. (1992) found that subjects recalled words from lists produced at a single speaking rate better than the same words produced at several different speaking rates. Interestingly, the differences appeared in the primacy portion of the serial position curve, suggesting greater difficulty in the transfer of items into long-term memory (Luce, Feustel, and Pisoni, 1983). Differences in speaking rate, like those observed for talker variability in our earlier experiments, suggest that perceptual encoding and rehearsal processes, which are typically thought to operate on only abstract symbolic representations, are also influenced by low-level perceptual sources of variability. If these sources of variability were filtered out or normalized by the perceptual system at relatively early stages of analysis, differences in recall performance would not be expected in memory tasks like the ones used in these experiments.

Taken together with the earlier results on talker variability, the findings on speaking rate suggest that details of the early perceptual analysis of spoken words are not lost as a result of perceptual analysis, but instead become an integral part of the neural representation of spoken words in memory. We have also found that in some cases increased stimulus variability in an experiment may actually help listeners to encode items into long-term memory (see Goldinger et al., 1991; Nygaard et al., 1992). Listeners encode speech signals in multiple ways along many perceptual dimensions and the human memory system apparently preserves these perceptual details much more precisely than researchers believed in the past.

## **EXPERIMENTS ON PERCEPTUAL LEARNING OF VOICES**

We have also been interested in perceptual learning, specifically the tuning or adaptation that occurs when a listener becomes familiar with the voice of a specific talker (Nygaard, Sommers, and Pisoni, 1994). This particular kind of perceptual learning has not received very much attention in the past despite the obvious relevance to problems of speaker normalization, acoustic-phonetic invariance, and the potential application to automatic speech recognition and speaker identification (Kakehi, 1992; Fowler, In Press). Our search of the research literature on talker adaptation revealed only a small number of studies on this topic, and all of them appeared in obscure technical reports from the mid 1950s. Thus, we decided to carry out a perceptual learning experiment of our own to see how knowledge of a talker's voice affects speech perception.

To determine how familiarity with a talker's voice affects the perception of spoken words, we had listeners learn to explicitly identify a set of unfamiliar voices over a nine-day period using a common set of names (i.e., Bill, Joe, Sue, Mary). After the subjects learned to recognize the voices explicitly, we presented them with a set of novel words mixed in noise at several signal-to-noise ratios; half the listeners heard the words produced by talkers that they were previously trained on (i.e., the familiar voices) and half the listeners heard the words produced by new talkers that they had not been exposed to previously (i.e., the novel voices). In this phase of the experiment, which was designed to measure speech intelligibility, subjects were now required to identify the words rather than explicitly recognize the voices as they had done in the earlier phase of the experiment.

The results of the speech intelligibility experiment are shown in Figure 7 for the two groups of subjects. We found that identification performance for the trained group was reliably better than the

control group at each of the signal-to-noise ratios tested. The subjects who had heard novel words produced by familiar voices were able to recognize words in noise more accurately than subjects who received the same novel words produced by unfamiliar voices. Two other groups of subjects were also run in the intelligibility experiment as controls; however, these subjects did not receive any training and were therefore not exposed to any of the voices prior to hearing the same set of words in noise. One control group received the set of test words presented to the trained experimental group; the other control group received the test words that were presented to the trained control subjects. The performance of both of the control groups was not only the same, but was equivalent to the intelligibility scores obtained by the trained control group. Only subjects in the experimental group who were explicitly trained on the voices showed the advantage in recognizing novel words produced by familiar talkers.

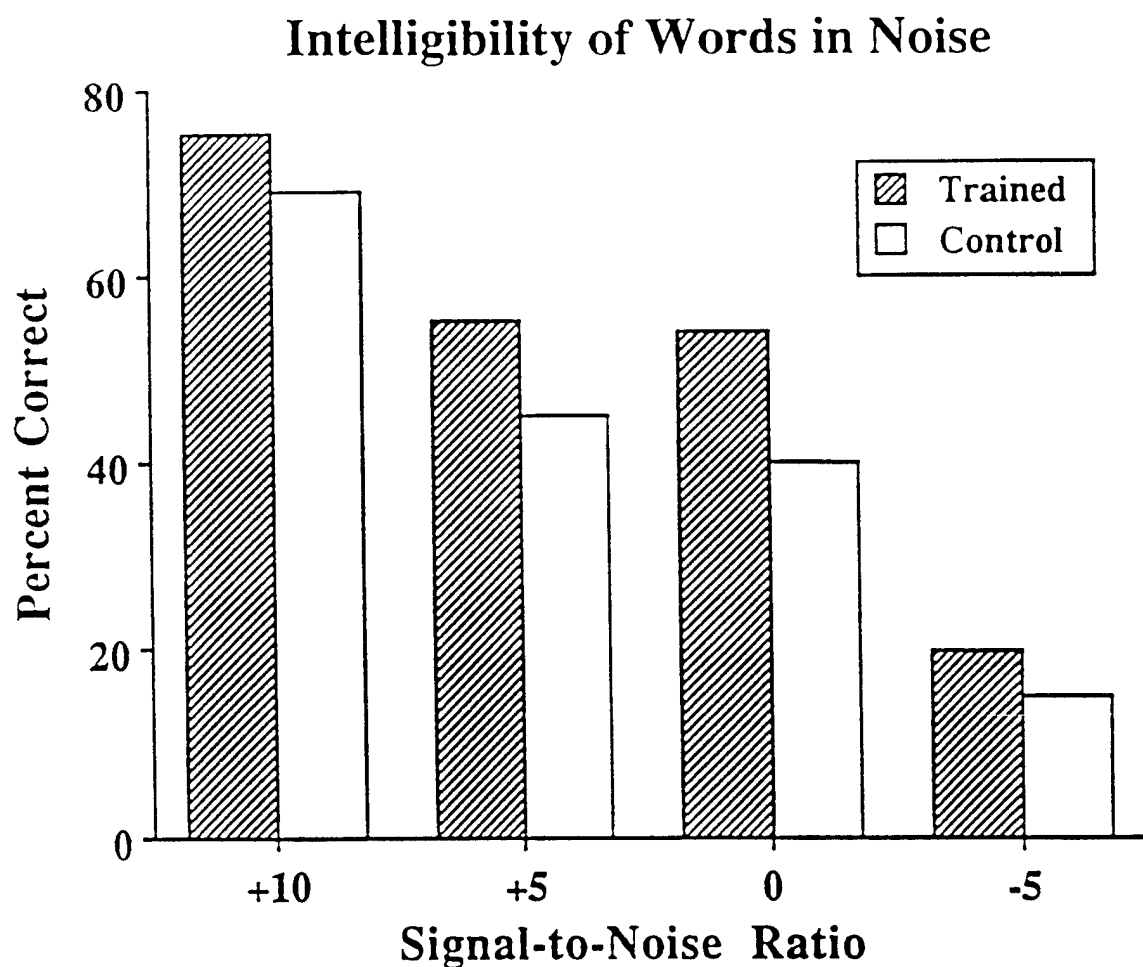


Figure 7. Mean intelligibility of words mixed in noise for trained and control subjects. Percent correct word recognition is plotted at each signal-to-noise ratio (from Nygaard et al., 1994).

The findings from this perceptual learning experiment demonstrate that exposure to a talker's voice facilitates subsequent perceptual processing of novel words produced by a familiar talker. Thus, speech perception and spoken word recognition draw on highly specific perceptual knowledge about a talker's voice that was obtained in an entirely different experimental task--explicit voice recognition as compared to a speech intelligibility test in which novel words were mixed in noise and subjects identified the items explicitly from an open response set.

What kind of perceptual knowledge does a listener acquire when he listens to a speaker's voice and is required to carry out an explicit name recognition task like our subjects did in this experiment? One possibility is that the analysis procedures or perceptual operations (Kolars, 1973) used to recognize the voices are retained in some type of procedural memory and these same processing routines are invoked again when the same voice is encountered in a subsequent intelligibility test. This kind of procedural knowledge might increase the efficiency of the perceptual analysis for novel words produced by familiar talkers because detailed analysis of the speaker's voice would not have to be carried out again. Another possibility is that specific instances--perceptual episodes or exemplars of each talker's voice--are stored in memory and then later retrieved during the process of word recognition when new tokens from a familiar talker are encountered (Jacoby and Brooks, 1984).

Whatever the exact nature of this information or knowledge turns out to be, the important point here is that prior exposure to a talker's voice facilitates subsequent recognition of novel words produced by the same talkers. Such findings demonstrate a form of implicit memory for a talker's voice that is distinct from the retention of the individual items used and the specific task that was employed to familiarize the listeners with the voices (Schacter, 1992; Roediger, 1990). These results provide additional support for the view that the neural representation of spoken words encompasses both a phonetic description of the utterance, as well as information about the structural description of the source characteristics of the specific talker. Thus, speech perception appears to be carried out in a talker-contingent manner; indexical and linguistic properties of the speech signal are apparently closely interrelated and are not dissociated in perceptual analysis as many researchers previously thought (see Nygaard et al., 1994). We believe these talker-contingent effects may provide a new way to deal with some of the old problems in speech perception that have been so difficult to resolve in the past.

## **ABSTRACTIONIST VS. EPISODIC APPROACHES TO SPEECH PERCEPTION**

The results we have obtained over the last few years raise a number of important questions about the theoretical assumptions that have been shared for many years by almost all researchers working in the field (Pisoni and Luce, 1986). Within cognitive psychology, the traditional approach to speech perception can be considered among the best examples of what have been called abstractionist accounts of categorization and memory (Jacoby and Brooks, 1984). Units of perceptual analysis in speech were assumed to be equivalent to the abstract idealized categories proposed by linguists in their formal analyses of language structure and function. The goal of speech perception studies was to find the physical invariants in the speech signal that mapped onto the symbolic phonetic categories of speech (Studdert-Kennedy, 1976). Emphasis was directed at separating stable, relevant features from the highly variable, irrelevant features of the signal. An important assumption of this traditional approach to perception and cognition was the process of abstraction and the reduction of information in the signal to a more efficient and economical symbolic code (Posner, 1969; Neisser, 1976).

Unfortunately, it became apparent very early in speech research that idealized linguistic units, such as phonemes or phoneme-like units, were highly dependent on the surrounding phonetic context and, moreover, that a wide variety of factors influenced their physical realization in the speech signal (Stevens, 1971; Klatt, 1986). Nevertheless, the search for acoustic invariance has continued in one way or another and still remains a central problem in the field today.

Recently, a number of studies on categorization and memory in cognitive psychology have provided evidence for the encoding and retention of episodic information and the details of perceptual analysis (Estes, 1993; Jacoby and Brooks, 1986; Brooks, 1978; Tulving and Schacter, 1990; Schacter, 1990). According to this approach, stimulus variability is considered to be "lawful" and "informative" to perceptual analysis (Elman and McClellan, 1986). Memory involves encoding specific instances, as well as the processing operations used in recognition (Kolers, 1973; Kolers, 1976b). The major emphasis of this view of perception and memory is on particulars, rather than abstract generalizations or symbolic coding of the stimulus input into idealized categories. Thus, the problems of variability and invariance found in speech perception can be studied in a different way by nonanalytic accounts of perception and memory with its emphasis on encoding of exemplars and specific instances of the stimulus environment rather than the search for physical invariants for abstract symbolic categories.

We believe that the findings from studies on nonanalytic cognition can be generalized to theoretical questions about the nature of perception and memory for speech signals and to assumptions about abstractionist representations based on formal linguistic analyses. When the criteria used for postulating nonanalytic representations are examined carefully, it immediately becomes clear that speech signals display a number of distinctive properties that make them especially good candidates for this approach to perception and memory (Jacoby and Brooks, 1984; Brooks, 1978). These criteria are summarized below and can be applied directly to speech perception and spoken language processing.

### **High Stimulus Variability**

Speech signals display a great deal of variability primarily because of factors related to the production of spoken language. Among these are within- and between-talker variability; changes in speaking rate and dialect; differences in social contexts; syntactic, semantic and pragmatic effects; as well as a wide variety of effects due to the ambient environment such as background noise, reverberation and microphone characteristics (Klatt, 1986). These diverse sources of variability consistently produce large changes in the acoustic-phonetic properties of speech and they need to be accommodated in theoretical accounts of speech perception.

### **Complex Category Relations**

The use of phonemes as perceptual categories in speech perception entails a set of complex assumptions about category membership which are based on formal linguistic criteria involving principles such as complementary distribution, free variation and phonetic similarity. The relationship between allophones and phonemes acknowledges explicitly the context-sensitive nature of the category relations that are used to define classes of speech sounds that function in similar ways in different phonetic environments.



## **Incomplete Information**

Spoken language is a highly redundant symbolic system which has evolved to maximize transmission of information. In speech perception, research has demonstrated the existence of multiple speech cues for almost every phonetic contrast. While these speech cues are, for the most part, highly context-dependent, they also provide partial information that can facilitate comprehension of the intended message when the signal is degraded. This feature of speech perception permits high rates of information transmission even under poor listening conditions.

## **High Analytic Difficulty**

Speech sounds are inherently multidimensional in nature. They encode a large number of quasi-independent articulatory attributes that are mapped on to the phonological categories of a specific language. Because of the complexity of speech categories and the high acoustic-phonetic variability, the category structure of speech is not amenable to simple hypothesis testing. As a consequence, it has been extremely difficult to formalize a set of explicit rules that can successfully map speech cues onto a set of idealized phoneme categories. Phoneme categories are also highly automatized. The category structure of a language is learned in a tacit and incidental way by young children. Because the criterial dimensional structures of speech are not typically available to consciousness, it has been difficult to make many aspects of speech perception explicit.

## **Three Domains of Speech**

Among category systems, speech appears to be unique in several respects because of the mapping between production and perception. Speech exists simultaneously in three very different domains: the acoustic domain, the articulatory domain and the perceptual domain. While the relations among these three domains is complex, they are not arbitrary because the sound contrasts used in a language function within a common linguistic signaling system that is assumed to encompass aspects of both production and perception. Thus, the phonetic distinctions generated in speech production by the vocal tract are precisely those same acoustic differences that are important in perceptual analysis (Stevens, 1972). Any theoretical account of speech perception must also take into consideration aspects of speech production and acoustics. The perceptual spaces mapped out in speech production have to be very closely correlated with the same ones used in speech perception. In learning the sound system of a language, the child must not only develop abilities to discriminate and identify sounds, but he/she must also be able to control the motor mechanisms used in articulation to generate precisely the same phonetic contrasts in speech production that he/she has become attuned to in perception. One reason that the developing perceptual system might preserve very fine phonetic details as well as characteristics of the talker's voice would be to allow a young child to accurately imitate and reproduce speech patterns heard in the surrounding language learning environment (Studdert-Kennedy, 1983). This skill would provide the child with an enormous benefit in acquiring the phonology of the local dialect from speakers he/she is exposed to early in life.

## GENERAL DISCUSSION

It has become common over the last 25 years to argue that speech perception is a highly unique process that requires specialized neural processing mechanisms to carry out perceptual analysis (Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967). These theoretical accounts of speech perception have typically emphasized the differences in perception between speech and other perceptual processes. Relatively few researchers working in the field of speech perception have tried to identify commonalities among other perceptual systems or draw parallels with speech perception. Our recent findings on the encoding of different sources of variability in speech and the role of long-term memory for specific instances are compatible with a rapidly growing body of research in cognitive psychology on implicit memory phenomena and nonanalytic modes of processing (Jacoby and Brooks, 1984; Brooks, 1978).

Traditional memory research has been concerned with explicit memory in which the subject is required to consciously access and manipulate recently presented information from memory using direct tests such as recall or recognition. This line of memory research has had a long history in experimental psychology and it is an area that most speech researchers are familiar with. In contrast, the recent literature on implicit memory phenomena has provided new evidence for unconscious aspects of perception, memory and cognition (Schacter, 1992; Roediger, 1990). Implicit memory refers to a form of memory that was acquired during a specific instance or episode and it is typically measured by indirect tests such as stem completion, cued recall, priming or changes in perceptual identification performance (Roediger, 1990; Roediger and McDermott, 1993). In these types of memory tests, subjects are not required to consciously recollect previously acquired information. In fact, in many cases, especially in processing spoken language, subjects may be unable to access the information deliberately or even bring it to consciousness (Studdert-Kennedy, 1974).

Studies of implicit memory have uncovered important new information about the effects of prior experience on perception and memory. In addition to traditional abstractionist modes of cognition which tend to emphasize symbolic coding of the stimulus input, recent experiments have provided evidence for a parallel nonanalytic memory system that preserves specific instances of stimulation as perceptual episodes or exemplars which are also stored in memory. These perceptual episodes have been shown to affect later processing activities. We believe that it is this implicit perceptual memory system that encodes the indexical information in speech about a talker's gender, dialect and speaking rate. And we believe that it is this memory system that encodes and preserves the perceptual operations or procedural knowledge that listeners acquire about specific voices that facilitates later recognition of novel words produced by familiar speakers.

Our findings demonstrating that spoken word recognition is talker-contingent and that familiar voices are encoded differently than novel voices raises a new set of questions concerning the long-standing dissociation between the linguistic properties of speech (the abstract, symbolic features, phonemes and words used to convey the linguistic message and the indexical properties of speech) and those personal or paralinguistic attributes of the speech signal which provide the listener with information about the form of the message (the speaker's gender, dialect, social class, and emotional state among other things). In the past, these two sources of information were separated for purposes

of linguistic analysis of the message. The present set of findings suggests this may have been an incorrect assumption.

Relative to the research carried out on the linguistic properties of speech, which has a history dating back to the late 1940s, much less is known about perception of the acoustic correlates of the indexical or paralinguistic functions of speech (Ladefoged, 1975; Laver and Trudgill, 1979). While there have been a number of recent studies on explicit voice recognition and identification by human listeners (Papcun, Kreiman, and Davis, 1989), very little research has been carried out on problems surrounding the implicit or unconscious encoding of attributes of voices and how this form of memory might affect the recognition process associated with the linguistic attributes of spoken words (Goldinger, 1992; Lively, 1994). A question that naturally arises in this context is whether or not familiar voices are processed differently than unfamiliar or novel voices. Perhaps familiar voices are simply recognized more efficiently than novel voices and are perceived in fundamentally the same way by the same neural mechanisms as unfamiliar voices? The available evidence in the literature has shown, however, that familiar and unfamiliar voices are processed differentially by the two hemispheres of the brain and that selective impairment resulting from brain language can affect the perception of familiar and novel voices in very different ways (see Kreiman and VanLancker, 1988; VanLancker, Cummings, Kreiman, and Dobkin, 1988; VanLancker, Kreiman, and Cummings, 1989).

Most researchers working in speech perception have adopted a common set of theoretical assumptions about the units of linguistic analysis and the goals of perceptual processing of speech signals. The primary objective was to extract the speaker's message from the acoustic waveform without regard to the source (Studdert-Kennedy, 1974). The present set of findings suggests that while the dissociation between indexical and linguistic properties of speech may have been a useful dichotomy for linguists who approach language as a highly abstract formalized symbolic system, the same set of assumptions may no longer be useful for speech scientists who are interested in describing and modeling how the human nervous system encodes speech signals and represents this information in long-term memory.

Our recent findings on variability suggest that fine phonetic details about the form of the signal are not lost as a consequence of perceptual analysis as widely assumed by researchers in the past. Attributes of the talker's voice are also not lost or normalized, at least not immediately after perceptual analysis has been completed. In contrast to the theoretical views that were very popular a few years ago, the present findings have raised some new questions about the problems of variability, invariance and perceptual normalization. For example, there is now sufficient evidence from perceptual experimentation to suggest that the fundamental perceptual categories of speech, i.e. phonemes and phoneme-like units, are probably not as rigidly fixed or well-defined physically as theorists once believed. These perceptual categories appear to be highly variable and their physical attributes have been shown to be strongly affected by a wide variety of contextual factors (Klatt, 1979). It seems very unlikely after some 45 years of research on speech that very simple physical invariants for phonemes will be uncovered from analysis of the speech signal. If invariants are uncovered, they will probably be very complex time-varying cues that are highly context-dependent.

Many of the theoretical views that speech researchers have held for a long time about language were motivated by linguistic considerations of speech as an idealized symbolic system essentially free

from physical variability. Indeed, variability in speech was considered by many researchers to be a source of noise--an undesirable set of perturbations on what was otherwise supposed to be an idealized sequence of abstract symbols arrayed linearly in time. Unfortunately, it has taken many years for speech researchers to realize that variability is an inherent characteristic of all biological systems including speech. Rather than view variability as noise, some theorists have recognized that variability might actually be useful and informative to human listeners who are able to encode speech signals in a variety of different ways depending upon the circumstances and demands of the listening task (Elman and McClellan, 1986). The recent proposals in the human memory literature for multiple memory systems suggest that the internal representation of speech is probably much more detailed and much more elaborate than previously believed from simply an abstractionist linguistic point of view. The traditional views about features, phonemes and acoustic-phonetic invariance are no longer adequate to accommodate the new findings that have been uncovered concerning context effects and variability in speech perception and spoken word recognition. In the future, it may be very useful to explore the parallels between similar perceptual systems such as face recognition and voice recognition. There is, in fact, some reason to suspect that parallel neural mechanisms may be employed in each case despite the obvious differences in modalities.

## CONCLUSIONS

The results summarized in this paper on the role of variability in speech perception are compatible with nonanalytic or instance-based views of cognition which emphasize the episodic encoding of specific details of the stimulus environment. Our studies on talker and rate variability and our new experiments on perceptual learning of novel voices have provided important information about speech perception and spoken word recognition and have served to raise a set of new theoretical questions for future research. In this section, I simply list the major conclusions.

First, our findings raise questions about previous views of the neural representation of speech. In particular, we have found that detailed instance-specific information about the source characteristics of a talker's voice are encoded into long-term memory. Whatever the internal representation of speech turns out to be, it is clear that it is not isomorphic with the linguist's description of speech as an abstract idealized sequence of segments. Mental representations of speech are much more detailed and more elaborate and they contain several sources of information about the talker's voice.

Second, our findings suggest a different approach to the problem of acoustic-phonetic variability in speech perception. Variability is not a source of noise; it is lawful and informative and provides potentially useful knowledge about the characteristics of a talker's voice and speaking rate as well as the phonetic context. These sources of information appear to be accessed when a listener hears novel words or sentences produced by a familiar talker. Variability provides important talker-specific information that affects encoding fluency and processing efficiency in a variety of tasks.

Third, our findings provide additional evidence that speech perception is highly sensitive to context and that details of the input signal are not lost or filtered out as a consequence of perceptual analysis. These results are consistent with recent proposals for the existence of multiple memory systems and the role of perceptual representation systems (PRS) in memory and learning. The present

findings also suggest a somewhat different view of the process of perceptual normalization which has generally focused on abstraction and stimulus reduction in categorization of speech sounds.

Finally, the results described here suggest several new directions for models of speech perception and spoken word recognition. These models are motivated by a different set of criteria than traditional abstractionist approaches to perception and memory. Exemplar-based or episodic models of categorization which emphasize instance-specific encoding provide a viable new theoretical alternative to the problems of invariance, variability and perceptual normalization that have been difficult to resolve with current models of speech perception that were inspired by formal linguistic analyses of language. We believe that many of the current theoretical problems in the field of speech perception can be approached in quite different ways when viewed within the general framework of nonanalytic or instance-based models of cognition which have alternative methods of dealing with the problems of stimulus variability, context effects and perceptual learning phenomena which have been the hallmarks of human speech perception for many years.

## ACKNOWLEDGMENTS

This research was supported by NIDCD Research Grant DC-00111-17, Indiana University in Bloomington. I thank Steve Goldinger, Scott Lively, Lynne Nygaard, Mitchell Sommers, Thomas Palmeri and John Karl for their help and collaboration in various phases of this research program.

## REFERENCES

- Brooks, L. "Nonanalytic Concept Formation and Memory for Instances," in E. Rosch and B. Lloyd (Eds.), *Cognition and Categorization*. Hillsdale NJ: Erlbaum (1978).
- Creelman, C.D. "Case of the unknown talker," *J. Acoust. Soc. Amer.*, 29, 655 (1957).
- Eich, J.E. "A composite holographic associative memory model," *Psychological Review*, 89, 627-661 (1982).
- Elman, J.L., and McClellan, J.L. "Exploiting Lawful Variability in the Speech Wave," *Invariance and Variability in Speech Processes*, Hillsdale NJ: Erlbaum, 360-380 (1986).
- Estes, W.K. *Classification and Cognition*, Oxford University Press: New York (1994).
- Fowler, C.A. "Listener-talker Attunements in Speech," in T. Tighe, B. Moore, and J. Santroch (Eds.), *Human Development and Communication Sciences*, Hillsdale NJ: Erlbaum (In press).
- Garner, W.R. *The processing of information and structure*, Potomac MD: Earlbau (1974).
- Goldinger, S.D. "Words and Voices: Implicit and Explicit Memory for Spoken Words," *Research on Speech Perception Technical Report No. 7*, Indiana University, Bloomington IN (1992).

Goldinger, S.D., Pisoni, D.B., and Logan, J.S. "On the Locus of Talker Variability Effects in Recall of Spoken Word Lists," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17, 152-162 (1991).

Hintzman, D.L. "Schema Abstraction in a multiple-trace memory model," *Psychological Review*, 93, 411-423, (1986).

Jacoby, L.L., and Brooks, L.R. "Nonanalytic Cognition: Memory, Perception, and Concept Learning," in G. Bower (Ed.), *The Psychology of Learning and Motivation*, New York: Academic Press, 1-47 (1984).

Takehi, K. "Adaptability to Differences between Talkers in Japanese Monosyllabic Perception," in Y. Tohkura, E. Vatikiotis-Bateson and Y. Sagisaka (Eds.), *Speech Perception, Production and Linguistic Structure*, Tokyo, Japan: IOS Press, Inc. (1992).

Karl, J.R., and Pisoni, D.B. "The role of talker-specific information in memory for spoken sentences," *J. Acoust. Soc. Amer.*, 95, 2873 (1994).

Klatt, D.H. "Speech Perception: A Model of Acoustic-Phonetic Analysis and Lexical Access," *Journal of Phonetics*, 7, 279-312 (1979).

Klatt, D.H. "The Problem of Variability in Speech Recognition and in Models of Speech Perception," in J.S. Perkell and D.H. Klatt (Eds.), *Invariance and Variability in Speech Processes*, Hillsdale NJ: Erlbaum (1986).

Kolers, P.A. "Remembering Operations," *Memory & Cognition*, 1, 347-355 (1973).

Kolers, P.A. "Pattern Analyzing Memory," *Science*, 191, 1280-1281 (1976b).

Kreiman, J. and VanLancker, D. "Hemispheric specialization for voice recognition: Evidence from dichotic listening," *Brain and Language*, 34, 246-252 (1988).

Ladefoged, P. *A Course in Phonetics*, New York: Harcourt Brace Jovanovich, Inc. (1975).

Laver, J. and Trudgill, P. "Phonetic and linguistic markers in speech," in K.R. Scherer and H. Giles (Eds.) *Social Markers in Speech*, Cambridge: Cambridge University Press, 1-31 (1979).

Liberman, A.M., Cooper, F.S., Shankweiler, D.P., and Studdert-Kennedy, M. "Perception of the Speech Code," *Psychological Review*, 74, 431-461 (1967).

Lively, S.E. "Preserving the perceptual record: Retention of voice information in long-term memory," *Research on Speech Perception Technical Report No. 9*, Indiana University, Bloomington IN (1994).

- Luce, P.A., Feustal, T.C., and Pisoni, D.B. "Capacity demands in short-term memory for synthetic and natural word lists, *Human Factors*, 25, 17-32 (1983).
- Martin, C.S., Mullennix, J.W., Pisoni, D.B., and Summers, W.V. "Effects of Talker Variability on Recall of Spoken Word Lists," *Journal of Experimental Psychology: Learning, Memory and Cognition*, 15, 676-684 (1989).
- Mullennix, J.W., and Pisoni, D.B. "Stimulus Variability and Processing Dependencies in Speech Perception," *Perception & Psychophysics*, 47, 379-390 (1990).
- Mullennix, J.W., Pisoni, D.B., and Martin, C.S. "Some Effects of Talker Variability on Spoken Word Recognition," *J. Acoust. Soc. Amer.*, 85, 365-378 (1989).
- Neisser, U. *Cognitive Psychology*, New York: Appleton-Century-Crofts (1976).
- Nygaard, L.C., Sommers, M.S., and Pisoni, D.B. "Effects of Speaking Rate and Talker Variability on the Recall of Spoken Words," *J. Acoust. Soc. Amer.*, 91, 2340 (1992).
- Nygaard, L.C., Sommers, M.S., and Pisoni, D.B. "Effects of speaking rate and talker variability on the representation of spoken words in memory," *Proceedings 1992 International Conference on Spoken Language Processing*, Banff, Canada, 12-17 October 1992.
- Nygaard, L.C., Sommers, M.S., and Pisoni, D.B. "Speech perception as a talker-contingent process," *Psychological Science*, 5, 42-46 (1994).
- Palmeri, T.J., Goldinger, S.D., and Pisoni, D.B. "Episodic Encoding of Voice Attributes and Recognition Memory for Spoken Words," *Journal of Experimental Psychology: Learning, Memory and Cognition*, 19, 1-20 (1993).
- Papcun, G., Kreiman, J., and Davis, A. "Long-term memory for unfamiliar voices," *J. Acoust. Soc. of Amer.*, 85, 913-925 (1989).
- Peters, R.W. "The Relative Intelligibility of Single-voice and Multiple-voice Messages Under Various Conditions of Noise," *Joint Project Report No. 56*, U.S. Naval School of Aviation Medicine, 1-9, Pensacola FL (1955).
- Pisoni, D.B. "Effects of Talker Variability on Speech Perception: Implications for Current Research and Theory," *Proceedings of 1990 International Conference on Spoken Language Processing*, Kobe, Japan, 1399-1407 (1990).
- Pisoni, D.B. "Some Comments on Talker Normalization in Speech Perception," in Y. Tohkura, E. Vatikiotis-Bateson and Y. Sagisaka (Eds.), *Speech Perception, Production and Linguistic Structure*, Tokyo, Japan: IOS Press, Inc. (1992a).

Pisoni, D.B. "Some comments on invariance, variability and perceptual normalization in speech perception," *Proc. 1992 Int'l Conf. on Spoken Lang. Process.*, Banff, Canada, 12-17 October 1992 (1992b).

Pisoni, D.B. "Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning," *Speech Communication*, 13, 109-125 (1993).

Pisoni, D.B., and Luce, P.A. "Speech Perception: Research, Theory, and the Principal Issues," in E.C. Schwab and H.C. Nusbaum (Eds.), *Pattern Recognition by Humans and Machines*, New York: Academic Press, 1-50 (1986).

Pisoni, D.B., Nusbaum, H.C., Luce, P.A., and Slowiaczek, L.M. "Speech perception, word recognition and the structure of the lexicon," *Speech Communication*, 4, 75-95 (1985).

Posner, M.I. "Abstraction and the process of recognition," in J.T. Spence and G.H. Bower (Eds.), *The Psychology of Learning and Motivation: Advances in Learning and Motivation*, New York: Academic Press (1969).

Roediger, H.L. "Implicit Memory: Retention Without Remembering," *American Psychologist*, 45, 1043-1056 (1990).

Roediger, H.L., and McDermott, K.B. "Implicit memory in normal human subjects" in F. Boller and J. Grafman (Eds.), *Handbook of Neuropsychology*, Elsevier Publishing: New York (1993).

Schacter, D.L. "Perceptual representation systems and implicit memory: Toward a resolution of the multiple memory systems debate," in A. Diamond (Ed.), *Development and Neural Basis of Higher Cognitive Function*, Annals of the New York Academy of Sciences, 608, 543-571 (1990).

Schacter, D.L. "Understanding Implicit Memory: A Cognitive Neuroscience Approach," *American Psychologist*, 47, 559-569 (1992).

Sommers, M.S., Nygaard, L.C., and Pisoni, D.B. "Stimulus variability and the perception of spoken words: Effects of variations in speaking rate and overall amplitude," *Proc. 1992 Int'l Conf. on Spoken Lang. Process.*, Banff, Canada, 12-17 October 1992.

Sommers, M.S., Nygaard, L.C., and Pisoni, D.B. "The Effects of Speaking Rate and Amplitude Variability on Perceptual Identification," *J. Acoust. Soc. Amer.*, 91, 2340 (1992).

Stevens, K.N. "Sources of Inter- and Intra-Speaker Variability in the Acoustic Properties of Speech Sounds," *Proceedings of the Seventh International Congress of Phonetic Sciences*, The Hague: Mouton (1971).

Stevens, K.N. "The quantal nature of speech: Evidence from articulatory acoustic data," in E.E. David, Jr. and P.B. Denes (Eds.), *Human communication: A unified view*, New York: McGraw-Hill (1972).



Studdert-Kennedy, M. "The Perception of Speech," in T.A. Sebeok (Ed.), *Current Trends in Linguistics*, The Hague: Mouton, 2349-2385 (1974).

Studdert-Kennedy, M. "Speech Perception," in N.J. Lass (Ed.), *Contemporary Issues in Experimental Phonetics*, New York: Academic Press (1976).

Studdert-Kennedy, M. "On learning to speak," *Human Neurobiology*, 2, 191-195 (1983).

Tulving, E., and Schacter, D.L. "Priming and Human Memory Systems," *Science*, 247, 301-306 (1990).

VanLancker, D.R., Cummings, J.L., Kreiman, J., and Dobkin, B. "Phonagnosia: A dissociation between familiar and unfamiliar voices," *Cortex*, 24, 195-209 (1988).

VanLancker, D.R., Kreiman, J., and Cummings, J. "Voice perception deficits: Neuroanatomical correlates of phonagnosia," *Journal of Clinical and Experimental Neuropsychology*, 11, 665-674 (1989).

# **SPEECH INTELLIGIBILITY EFFECTS IN A DUAL-TASK ENVIRONMENT**

**David G. Payne and Michael J. Wenger**

**Department of Psychology and  
Center for Cognitive and Psycholinguistic Sciences  
State University of New York at Binghamton**

***Abstract.*** Four experiments assessed whether changes in the level of speech intelligibility in an auditory task impact performance in concurrently performed visual tasks. Each experiment used an auditory memory search task in which subjects memorized a set of words and then decided whether auditorily presented probe items were members of the memorized set. The visual tasks used were an unstable tracking task, a spatial decision-making task, a mathematical reasoning task, and a probability monitoring task. Results showed that performance on the unstable tracking and probability monitoring tasks were unaffected by the level of speech intelligibility on the auditory task, whereas accuracy in the spatial decision-making and mathematical processing tasks was significantly worse at low speech intelligibility levels. The findings are interpreted within the framework of multiple resource theory.

## **INTRODUCTION**

It is well documented that a wide variety of factors affect speech intelligibility, and a great deal of research has aimed at understanding how changes in speech intelligibility affect performance in auditory tasks. There is also considerable evidence that changing task demands in an auditory task (e.g., by varying talker variability or utilizing either natural or synthetic speech) can exert very specific effects on the types of cognitive resources required to perform the task (e.g., Luce, Fuestle, and Pisoni, 1983; Goldinger, Pisoni, and Logan, 1991; Martin, Mullennix, Pisoni, and Summers, 1989). One question that has received far less attention is whether changes in speech intelligibility have the potential to affect performance in other, nonauditory, tasks. If decreasing speech intelligibility levels results in an increase in the perceptual/cognitive resources required to process the auditory signal, then it seems likely that these increases could also affect performance in concurrent tasks. The four experiments reported here were designed to test whether changes in speech intelligibility might affect performance levels in nonauditory tasks and, if so, what type(s) of tasks might be affected.

It seems reasonable to assume that in multitask environments changes in the demands of one or more tasks might affect performance in the remaining tasks. The more useful question, however, is whether there is any principled means by which to predict when changes in speech intelligibility level might affect performance in other tasks. This latter question motivated the

present research. A primary objective of these experiments was to, on the basis of existing and tested theory, identify the types of concurrent nonauditory tasks likely to be affected by changes in speech intelligibility.

The present research was conducted within the theoretical framework provided by multiple resource theory (e.g., Navon and Gopher, 1979; Wickens, 1980, 1984). Navon and Gopher (1979) proposed that the human cognitive system could be viewed as comprised of a limited number of processing resources. These processing resources are hypothetical constructs that refer to underlying commodities that enable a person to perform some task(s). According to this framework, resources are limited in the sense that specific resources may only be allocated to specific processes or subprocesses. There is ample evidence to support the general claim that some tasks may be performed simultaneously with little change in performance levels (e.g., Allport, Antonis, and Reynolds, 1972; Shaffer, 1975), whereas other tasks will greatly interfere with each other (e.g., Brooks, 1968).

Wickens (1980, 1984) identified the following as reasonable candidates for processing resources: (a) the type of input and output modality (e.g., visual vs. auditory stimuli; manual vs. vocal responses); (b) the code or representational format utilized by the subject (e.g., a linguistic vs. a spatial code), and (c) the stage of processing (e.g., encoding, central processing, and response selection/execution). For present purposes the predictions derived from Wickens' multiple resource theory are straightforward. If a visual and an auditory task employ the same resources, then performing these tasks together should be quite difficult and we should see performance decrements, especially as we increase the difficulty levels or resource demands of the tasks of interest. If, however, the two tasks tap different resources, then subjects should be able to perform the two tasks together efficiently and there should be no effect on concurrent task performance when the difficulty level of one task is increased. We assume here that changes in speech intelligibility levels increase the difficulty level of auditory tasks by increasing the amount of central processing resources that are required to analyze the auditory stimulus and use the stimulus information to perform the task of interest.

To test the multiple resource theory predictions in the present study, we employed a dual-task methodology (cf., Ogden, Levine, and Eisner, 1979) in which subjects are required to perform two tasks, singly and in conjunction. The single-task conditions provide baseline performance levels, and the dual-task conditions allow us to determine whether the two tasks selectively interfere with one another. Finally, task difficulty level is manipulated in both tasks so as to vary the amount of resources allocated to the tasks, thereby allowing us to look for the presence or absence of selective interference effects (e.g., the effect of changing speech intelligibility on performance of a visual task).

In order to test the effects of changes in speech intelligibility on concurrent visual task performance within the context of multiple resource theory, it is necessary to (a) have a reliable method for establishing the desired speech intelligibility levels, and (b) select visual tasks that selectively tap different cognitive resources. Fortunately, both of these needs can be met by using procedures validated in previous research.

Peters and Garinther (1990; see also Whitaker, Peters, and Garinther, 1989, 1990) have employed a chopping circuit designed by the U. S. Army Human Engineering Laboratory to vary the intelligibility levels in auditory tasks. (This chopping circuit will be described in detail in the Method section of Experiment 1.) Whitaker, Peters and Garinther's research is important to the present study because their results indicated that the chopping circuit parameters can be easily adjusted so as to produce desired levels of speech intelligibility.

The visual tasks used to tap into the resources identified by Wickens' (1980, 1984) multiple resource theory were taken from the Criterion Task Set (CTS) developed by Shingledecker, Acton, and Crabtree (1982). The CTS was originally developed within the framework of multiple resource theory and has undergone extensive validation (see Schlegel and Shingledecker, 1985). In the present research we utilized four of the CTS tasks, two that rely on visual encoding and manual responding (unstable tracking and probability monitoring) and two that require the central processes of working memory and decision-making (spatial processing and math processing). These tasks were selected for two reasons. First, they correspond quite closely to cognitive resources identified in Wickens' (1980, 1984) multiple resource model. Second, they are analogous to real-world perceptual/motor tracking tasks (e.g., sighting on a moving target, driving) and cognitive/decision-making tasks (e.g., distinguishing friend from foe).

The auditory task used in the present experiments was an auditory analog of the memory search task developed by Sternberg (1969). In this task subjects memorize a small set of spoken words and are then presented with a series of probe items, some of which came from the memorized target set. The subjects' task is to decide as quickly as possible if each probe item is a member of the target set. The primary performance measure is subjects' reaction time (RT), and task difficulty is manipulated by varying the number of items in the target set.

## **EXPERIMENT 1**

Experiment 1 was designed to test whether visual task performance levels would be affected by changes in speech intelligibility in a concurrent auditory task when the visual and auditory tasks tapped into different mental resources. The visual task was the CTS unstable tracking task, which is presumed to load heavily on visual encoding and manual responding. The auditory task was the auditory Sternberg Task performed at four different levels of speech intelligibility. Note that the Sternberg Task requires subjects to (a) maintain items in short-term memory and (b) make decisions concerning a series of probe items. Since both memory and decision-making are central processes, it seems reasonable to assume that changes in intelligibility levels will affect central resources.

If unstable tracking relies predominantly on visual encoding and manual responding, with relatively little central processing required, then according to Wickens' (1980, 1984) multiple resource framework, there should be no effect of speech intelligibility on tracking task performance. There should, however, be effects of the two difficulty manipulations on the corresponding tasks, and speech intelligibility should affect performance in the auditory Sternberg Task.

## **Method**

### **SUBJECTS**

Twenty-eight students enrolled in Introductory Psychology at SUNY-Binghamton participated in Experiment 1. Subjects were tested individually in a small (1.8 m x 3.0 m) sound-attenuated room in a single experimental session lasting approximately two hours.

### **DESIGN**

The experiment involved one between-subjects variable--level of difficulty (Easy vs. Difficult) on the unstable tracking task. The three within-subjects variables were (a) level of speech intelligibility (20%, 40%, 60%, 80%), (b) number of target items (2, 4) in the memory set on the auditory Sternberg Task, and (c) trial type, either single-task (unstable tracking, auditory Sternberg) or dual-task (unstable tracking performed concurrently with the auditory Sternberg Task).

### **APPARATUS**

The unstable tracking task was controlled by a Commodore 64 microcomputer interfaced with a Commodore Model 1702 color monitor and a 1.25 in. (3.5 cm) diameter rotary knob mounted in a 4 in. x 2 in. (10 x 5 cm) response box. Subjects used their right hand when performing the unstable tracking task.

The auditory stimuli were recorded using a high-quality microphone and a Data Translation analog-to-digital interface card (Model DT2801) along with an IBM compatible microcomputer equipped with an 80286 microprocessor operating at 12.5 Mhz. The stimuli were presented to subjects using the Data Translation card's digital-to-analog capabilities. These items were processed and speech intelligibility levels were varied by the chopping circuit described by Peters and Garinther (1990), amplified using a Radio Shack Model SA-150 amplifier, and presented to subjects over Realistic stereo headphones (Model Nova 65).

The chopping circuit used in the present experiments removes portions of the speech signal by chopping the signal for varying durations. The chopping circuit gated the speech signal at 60 Hz with a duty cycle variable from 0% to 95%. The circuit also adds a speech-shaped masking noise (i.e., pink noise) passed through a first-order, low-pass filter of 250 Hz and a first-order, high-pass filter of 350 Hz. To maintain comparability of the current results with those of previous studies (Peters and Garinther, 1990; Whitaker et al., 1990), masking noise was employed in the present study.

### **MATERIALS**

The target and nontarget items for the auditory Sternberg Task and the to-be-identified items for the Modified Rhyme Test (MRT) trials consisted of six lists of 50 items each developed by House, Williams, Hecker, and Kryter (1956). Items from List A were used as target and

nontarget items in the auditory Sternberg Task. The remaining five lists were used in the MRT trials. All stimuli were spoken by a male native English speaker and were digitally recorded (12 bit, 10 KHz sample rate).

### MODIFIED RHYME TEST

Subjects were presented with 50 words, one at a time. Each target word was preceded by a carrier phrase ("The next word is . . ."). After each word was presented, six alternative targets were presented on a CRT and the subject indicated which word they thought had just been presented auditorily by pressing a number (1-6) on the numeric key pad. After each response was made, the CRT screen was cleared and 2 secs later the carrier phrase and the next target item were presented. Speech intelligibility level obtained on each trial was operationally defined as the percentage of correct responses on that list.

### AUDITORY STERNBERG TASK

Subjects were presented with either two or four items to memorize. The items selected as target (and nontarget) items on each trial were randomly determined and the random ordering was held constant for all subjects. After the target items were presented, subjects could elect to review the items if they wished; otherwise, subjects signaled the experimenter and the trial was begun.

Each trial consisted of the presentation of an equal number of target and nontarget items, presented in a random order. Subjects were instructed to use their left hand to press the 1 key on a numeric key pad if the item came from the set of memorized target items for that trial (positive probes) and press the 3 key when the item was not a member of the memorized set. Subjects were instructed to respond as quickly and accurately as possible. Trials lasted approximately three minutes.

### VISUAL TASK

The visual task used in Experiment 1 was the CTS unstable tracking task from Shingledecker et al. (1982). The CTS unstable tracking task is designed to place variable demands on human information processing resources involving the execution of rapid and accurate manual responses. Subjects view a video screen displaying a fixed target area centered on the screen. A cursor moves vertically from the center of the screen, and the operator attempts to keep the cursor centered over the target area by rotary movements of the control knob. The system represented by the task is an inherently unstable one, and the dynamics of the task are a first-order divergent element of the following form:

$$P(s) = [\lambda / (s - \lambda)]e^{-\lambda s} \quad (1)$$

where  $\lambda$  (lambda) is selected by the experimenter to vary the manual control workload. The operator's input introduces error which is magnified by the system with the result that it becomes

increasingly necessary to respond to the velocity of the cursor movement as well as cursor position. No external forcing function is applied to the tracking loop. The unstable dynamics are simply excited by human tracking remnant and by noise in the controller digitization process. If the subject loses control and the cursor reaches the edge of the display, it is automatically reset to display center and the subject continues tracking. Subjects performed the tracking task continuously for three minutes using their right hand.

## **Procedure**

All subjects completed four blocks of trials, one at each level of speech intelligibility. Order of speech intelligibility level was varied using a balanced Latin square design. Within each block of trials, subjects completed seven tasks in the following order: (a) initial MRT trial, (b) single-task auditory Sternberg with two targets, (c) single-task unstable tracking (at the level of difficulty appropriate for that subject), (d) dual-task trial (auditory Sternberg + unstable tracking) with two target items in the Sternberg Task, (e) single-task auditory Sternberg with four target items, (f) dual-task trial with four target items in the Sternberg Task, and (g) final MRT trial.

Each block of critical trials was identical with the exceptions that (a) a different level of speech intelligibility was used on each block, (b) different lists were used on each MRT trial, and (c) different targets and nontargets were used on each Sternberg Task trial. Order of presentation of the MRT lists was counterbalanced across subjects and across levels of speech intelligibility. In the auditory Sternberg Task, assignment of items to the positive set (i.e., memory set) and negative set (i.e., nontargets) was also counterbalanced across subjects and speech intelligibility levels.

Based on previous research by Peters and Garinther (1990) and pilot research in the senior author's laboratory, the duty cycle setting on the chopping circuit for the 80%, 60%, 40%, and 20% speech intelligibility conditions resulted in speech signals that produced response accuracy levels of 83%, 62%, 31%, and 10.3%. Following this initial block of MRT trials, subjects performed the remaining six tasks in the order described above. Subjects completed the four blocks of trials in order, with a brief rest being given between blocks 2 and 3.

## **Results**

Several dependent variables were used to assess performance levels in the single and dual-task conditions. Nominal speech intelligibility levels were measured by computing the mean percentage of correct responses on the initial and final MRT given within each block of trials. For the auditory Sternberg Task, accuracy and reaction time were the dependent variables of interest, while for the unstable tracking task we used the mean absolute tracking error (in pixels). Note that for the unstable tracking task lower average absolute tracking error scores correspond to better performance levels.

## SPEECH INTELLIGIBILITY

Presented in the upper panel of Table 1 are the mean intelligibility levels (calculated as the mean of the first and second MRT) for each intelligibility level condition. (In this and each of the following experiments, initial analyses indicated that there were no effects of MRT list [ $p > 0.20$ ] in either the RT or accuracy measures and hence no further mention will be made of this factor.) As the data in Table 1 indicate, the observed intelligibility levels were close to the desired intelligibility levels.

Also presented in Table 1 are the MRT scores from Experiments 2-4. These experiments used procedures similar to those of Experiment 1 and the data are consistent with those from Experiment 1. Statistical analyses of these data yielded exactly the same results as in Experiment 1. Since the MRT data are not of primary concern here (other than to demonstrate that we succeeded in varying intelligibility levels), no further mention will be made of these data.

A 2 (Unstable Tracking Difficulty: Easy vs. Difficult) x 4 (Intelligibility Level: 80%, 60%, 40%, 20%) mixed-factor analysis of variance (ANOVA) performed on these data indicated a significant effect of Intelligibility Level,  $F(3,78) = 407.1$ ,  $p < 0.05$ . (Unless otherwise noted, all effects reported as significant had  $p < 0.05$ .) Newman-Keuls pairwise comparisons indicated that the observed intelligibility level in each condition differed significantly from all other conditions. Neither the main effect of Unstable Tracking Difficulty Level nor the Unstable Tracking Difficulty x Intelligibility Level interaction was significant.

## SINGLE-TASK TRIALS

Auditory Sternberg. Performance on the single-task auditory Sternberg trials was examined by comparing recognition accuracy and response latency (see Table 2). These data were analyzed using separate 2 (Tracking Task Difficulty: Easy vs. Difficult) x 2 (Number of Sternberg Target Items: 2 vs. 4) x 4 (Intelligibility Level: 80%, 60%, 40%, 20%) mixed-factor ANOVAs, one for each performance measure. These two analyses produced completely parallel results. As expected, performance levels were affected by the level of intelligibility and by the number of target items. There were significant main effects of the level of intelligibility,  $F(3,78) = 200.8$ , and  $F(3,78) = 38.2$ , and the number of target items,  $F(1,26) = 49.7$ , and  $F(1,26) = 35.2$ , for both the accuracy and RT measures, respectively. There was no effect of Tracking Task Difficulty for either the accuracy measure,  $F(1,26) = 2.00$ , or the RT measure  $F(1,26) = 2.45$ .

Unstable Tracking. Presented in Figure 1 are the mean error scores for the single- and dual-task trials in the easy and difficult unstable tracking conditions. (Note that for the single-task trials intelligibility level corresponds to a pseudo-variable, i.e., the intelligibility level for the auditory Sternberg Task in the block of trials in which the single-task tracking data were obtained. This fact holds for the single-task visual trials in Experiments 2-4 as well.)

Replicating previous studies (e.g., Schlegel and Shingledecker, 1985), for the single-task trials there were large differences in the performance levels for the easy and difficult tracking tasks. Furthermore, there was no effect of speech intelligibility for the block of trials within



Table 1

Experiments 1-4: Mean Percentage Correct Responses on the Modified Rhyme Test (MRT) for the Four Intelligibility Conditions

Experiment 1: Unstable Tracking				
<u>Difficulty Level</u>	<u>Intelligibility Condition</u>			
	<u>80%</u>	<u>60%</u>	<u>40%</u>	<u>20%</u>
Easy	78	70	50	26
Difficult	71	61	46	23
Experiment 2: Spatial Processing				
<u>Difficulty Level</u>	<u>Intelligibility Condition</u>			
	<u>80%</u>	<u>60%</u>	<u>40%</u>	<u>20%</u>
Easy	75	63	47	25
Difficult	73	66	43	22
Experiment 3: Mathematical Reasoning				
<u>Difficulty Level</u>	<u>Intelligibility Condition</u>			
	<u>80%</u>	<u>60%</u>	<u>40%</u>	<u>20%</u>
Easy	74	66	41	23
Difficult	73	67	43	25
Experiment 4: Probability Monitoring				
<u>Difficulty Level</u>	<u>80%</u>	<u>60%</u>	<u>40%</u>	<u>20%</u>
Easy	75	70	44	21
Medium	73	67	46	23
Difficult	73	66	52	24

**Table 2**

**Experiment 1: Mean Percent Correct Responses and Reaction Times (sec) for the Single and dual-task Auditory Sternberg Trials**

<u>Difficulty Level of Visual Task/</u>	<u>Intelligibility Condition</u>			
<u>Number of Targets</u>	<u>80%</u>	<u>60%</u>	<u>40%</u>	<u>20%</u>
<u>Percent Correct: single-task Trials</u>				
Easy - 2 Targets	99	99	96	60
Difficult - 2 Targets	99	98	87	60
Easy - 4 Targets	97	87	85	57
Difficult - 4 Targets	93	90	78	54
<u>Mean: 2 Targets</u>	99	98	92	60
<u>Mean: 4 Targets</u>	95	89	82	56
<u>Percent Correct: dual-task Trials</u>				
Easy - 2 Targets	98	97	92	62
Difficult - 2 Targets	93	91	86	56
Easy - 4 Targets	96	97	86	60
Difficult - 4 Targets	90	91	77	56
<u>Mean: 2 Targets</u>	95	94	89	60
<u>Mean: 4 Targets</u>	93	94	82	58

(Table 2, continued)

<u>Mean Reaction Time: single-task Trials</u>				
Easy - 2 Targets	1.125	1.115	1.214	1.565
Difficult - 2 Targets	1.172	1.214	1.440	1.582
Easy - 4 Targets	1.279	1.284	1.424	1.591
Difficult - 4 Targets	1.273	1.456	1.563	1.726
<u>Mean: 2 Targets</u>	1.148	1.164	1.327	1.573
<u>Mean: 4 Targets</u>	1.225	1.249	1.432	1.586

---

<u>Mean Reaction Time: dual-task Trials</u>				
Easy - 2 Targets	1.186	1.212	1.254	1.556
Difficult - 2 Targets	1.153	1.276	1.222	1.514
Easy - 4 Targets	1.233	1.235	1.355	1.526
Difficult - 4 Targets	1.258	1.338	1.369	1.458
<u>Mean: 2 Targets</u>	1.169	1.244	1.238	1.535
<u>Mean: 4 Targets</u>	1.245	1.287	1.362	1.492

---

which this single-task trial was run. The lack of an effect of speech intelligibility is, of course, not surprising in that subjects were not performing an auditory task during the single-task unstable tracking task trial.

These observations concerning the single-task trials were supported by the results of a 2 (Tracking Task Difficulty: Easy vs. Difficult) x 4 (Speech Intelligibility Levels: 80%, 60%, 40%, 20%) mixed-factor ANOVA. There was a significant main effect of Tracking Task Difficulty,  $F(1,26) = 143.2$ . Neither the main effect of Speech Intelligibility nor the Tracking Task Difficulty x Speech Intelligibility interaction was significant ( $F$ 's < 1.0).

#### DUAL-TASK TRIALS

Auditory Sternberg. The accuracy and response latency data from the auditory Sternberg Task in the dual-task trials were analyzed using separate 2 (Number of Target Items: 2 vs. 4) x 4 (Intelligibility: 80%, 60%, 40%, 20%) x 2 (Tracking Task Difficulty: Easy vs. Hard)

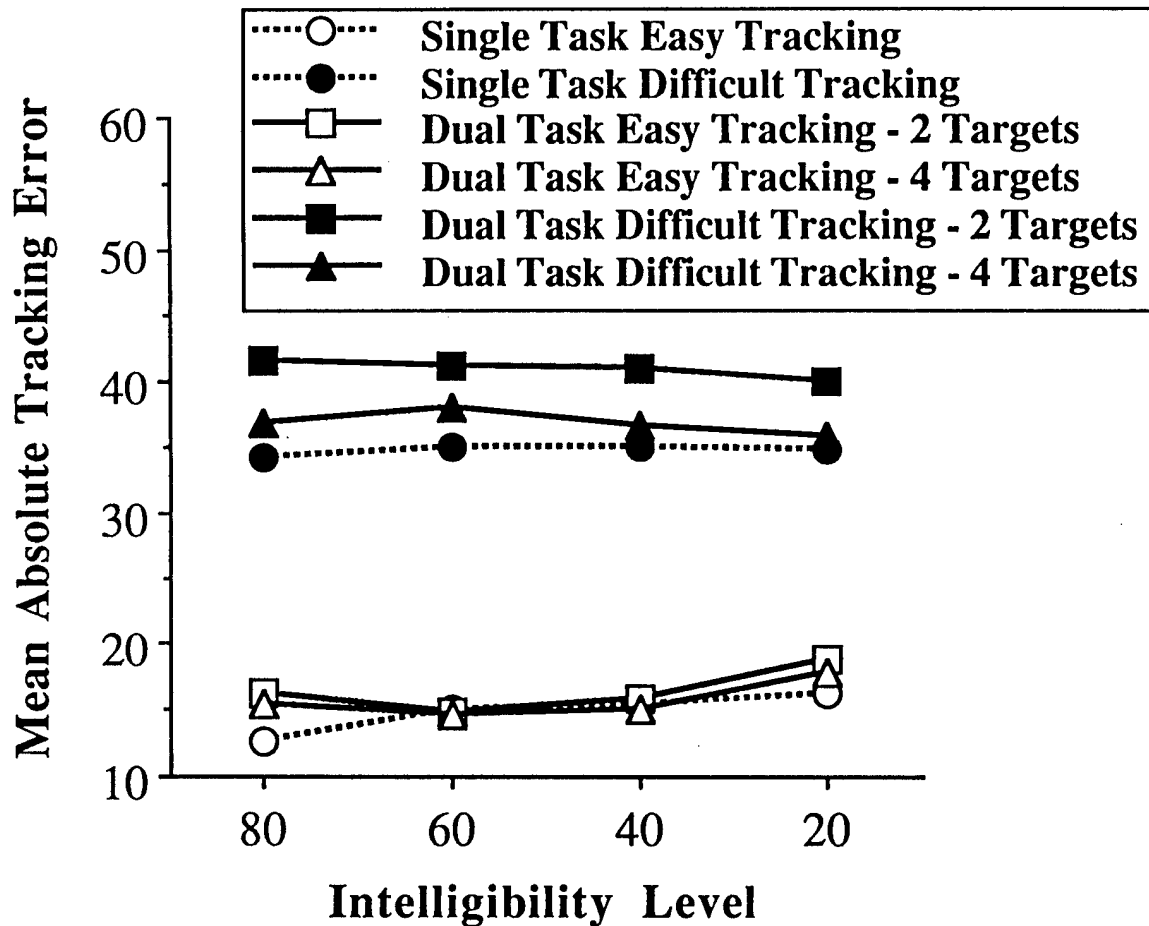


Figure 1. Experiment 1: Performance levels in the unstable tracking task for the easy and difficult unstable tracking conditions in the single-task and dual-task trials.

mixed-factor ANOVAs, one for each dependent variable. Presented in Table 2 are the mean percent correct responses and mean RTs for the 2- and 4-alternative auditory Sternberg trials from the easy and difficult dual-task trials conducted at the four intelligibility levels. As with the single-task trials, speech intelligibility level produced a significant effect on both response accuracy,  $F(3,78) = 158.2$ , and response latency,  $F(3,78) = 14.9$ . The number of target items also affected both accuracy,  $F(1,26) = 5.3$ , and RT,  $F(1,26) = 7.6$ , indicating that subjects were faster and more accurate in the two-target condition than the four-target condition.

Results also indicated that for the accuracy measure there was a significant main effect of visual task difficulty. When performing the auditory Sternberg Task in conjunction with the easy unstable tracking task, subjects made 85.9% correct responses in contrast to 79.9% correct with the hard version of the tracking task,  $F(1,26) = 12.8$ . However, the difficulty of the unstable tracking task did not interact with speech intelligibility level,  $F < 1.0$ .

Overall, then, the results from the auditory Sternberg dual-task condition indicate that performance was impacted by the number of target items and by the level of speech intelligibility. The main effect of unstable tracking task difficulty on the reaction time measure indicates that the tracking task difficulty did vary the overall joint task difficulty. Importantly, in the auditory Sternberg Task the tracking task difficulty did not interact with speech intelligibility level, suggesting that either (a) the tracking task and the auditory Sternberg Task tapped into different resources, or (b) the two tasks tapped into the same resources, but the resulting performance decrement was reflected only in the unstable tracking task.

Unstable Tracking. Presented in Figure 1 are the mean absolute tracking error scores for the easy and difficult unstable tracking tasks for the dual-task trials for each of the four intelligibility conditions. Replicating the single-task data, there was a significant effect of tracking task difficulty,  $F(1,26) = 156.8$ . More importantly, however, there was no evidence of a main effect of speech intelligibility, nor did intelligibility interact with any other variable (all  $F$ 's  $< 1.0$ ).

Analyses of these data revealed that there was a significant main effect of the number of target items in the auditory Sternberg Task,  $F(1,26) = 8.5$ , as well as a significant Tracking Task Difficulty x Number of Target Items interaction,  $F(1,26) = 4.3$ . The source of these effects lies primarily in the difficult tracking task condition; in the easy tracking task condition there was a nonsignificant difference between the two- and four-target item conditions (16.4 vs. 15.7),  $F < 1.0$ , while there was a significant difference in the difficult tracking task (41.1 vs. 36.8),  $F(1,26) = 12.5$ . Thus, subjects performed more poorly on the tracking task when there were two target items on the Sternberg Task than when there were four items. This difference may reflect a trade-off in task emphasis in that as the Sternberg Task increased in difficulty, subjects elected to focus more on the tracking task. Note, however, that there is no evidence of such a change in task emphasis as a function of speech intelligibility level. This is consistent with the hypothesis that speech intelligibility impacts resources not shared by the unstable tracking task.

## EXPERIMENT 2

Experiment 2 was designed to test whether a visual task that places large demands on central processing resources involved in performing the auditory Sternberg Task would be affected by changes in speech intelligibility. Toward this goal we conducted a replication and extension of Experiment 1, this time using the spatial processing task from the CTS. In the spatial processing task, subjects are presented with pairs of histograms, the first presented in a vertical orientation and the second rotated at 0, 90, 180, or 270 degrees from the original. The subjects' task is to decide if the two histograms are identical in overall shape, regardless of the orientation of the second (comparison) stimulus.

## Method

### SUBJECTS, DESIGN, AND APPARATUS

Twenty-eight subjects were recruited from the same source as in Experiment 1. The experimental design and apparatus were the same as in Experiment 1 with one exception. The visual task employed in Experiment 2 was the CTS spatial processing task, in which subjects viewed computer-generated pairs of histograms presented on the screen. Each histogram bar could assume any of six arbitrary heights; the first histogram in each pair appeared in the vertical orientation and was labeled with a 1. The second histogram was presented at either the same vertical orientation, rotated 90 degrees left or right, or rotated 180 degrees, and was labeled with a 2. In the easy version of the spatial processing task, each histogram contained four bars, while in the difficult condition there were six bars. The first (target) stimulus was presented for 3.0 secs followed by a short pause. The comparison stimulus was removed as soon as the subject made a response, and the screen remained blank until the next target stimulus was presented.

## Results

### SINGLE-TASK TRIALS

Auditory Sternberg. The mean percent correct responses and RTs for the two- and four-alternative single-task Sternberg trials are presented in Table 3. Replicating Experiment 1, both accuracy and RT were affected by the level of speech intelligibility,  $F(3,78) = 103.8$ , and  $F(3,78) = 13.4$ , respectively. The number of target items affected subjects' accuracy, with 83.4% correct responses in the two-alternative condition and 78.6% correct responses in the four-alternative condition,  $F(1,26) = 8.2$ .

Spatial Processing. Presented in Figure 2 are the mean percent correct responses (top panel) and mean RTs (bottom panel) in the single- and dual-task spatial processing trials. The single-task data indicate that in the single-task trials subjects were more accurate in the easy spatial processing condition than in the difficult condition,  $F(1,26) = 11.0$ . In addition, speech intelligibility also affected performance levels,  $F(3,78) = 2.7$ . Although the Spatial Processing Difficulty level x Speech Intelligibility interaction was not significant,  $F < 1.0$ , visual inspection of the data suggests that the difficult condition was more affected by changes in speech intelligibility than was the easy condition. Simple effects tests revealed that there was a significant effect of speech intelligibility for the difficult spatial processing task but not for the easy spatial processing task.

Response latency in the spatial processing task was slightly, but not significantly, ( $p = 0.26$ ), longer in the difficult spatial processing condition than in the easy spatial processing condition (1.099 vs. 1.004 secs, respectively). There was no evidence of speech intelligibility affecting RTs in the spatial processing tasks, with mean RTs of 1.063, 1.071, 1.040, and 1.034 secs for the 20%, 40%, 60%, and 80% speech intelligibility conditions, respectively.

Table 3

Experiment 2: Mean Percent Correct Responses and Reaction Times (sec) for the Single and dual-task Auditory Sternberg Trials

<u>Difficulty Level of Visual Task/</u>	<u>Intelligibility Condition</u>			
<u>Number of Targets</u>	<u>80%</u>	<u>60%</u>	<u>40%</u>	<u>20%</u>
<u>Percent Correct: single-task Trials</u>				
Easy - 2 Targets	97	97	88	58
Difficult - 2 Targets	94	96	76	61
Easy - 4 Targets	91	90	77	55
Difficult - 4 Targets	92	91	79	54
<u>Mean: 2 Targets</u>	95	96	82	60
<u>Mean: 4 Targets</u>	91	90	78	55
<u>Percent Correct: dual-task Trials</u>				
Easy - 2 Targets	94	97	83	59
Difficult - 2 Targets	92	94	83	57
Easy - 4 Targets	92	85	77	56
Difficult - 4 Targets	90	87	76	52
<u>Mean: 2 Targets</u>	93	95	83	58
<u>Mean: 4 Targets</u>	91	86	76	54

(Table 3, continued)

Mean Reaction Time: single-task Trials

Easy - 2 Targets	1.175	1.227	1.294	1.614
Difficult - 2 Targets	1.082	1.211	1.760	1.685
Easy - 4 Targets	1.241	1.323	1.419	1.690
Difficult - 4 Targets	1.284	1.380	1.442	1.491
<u>Mean: 2 Targets</u>	1.129	1.219	1.527	1.650
<u>Mean: 4 Targets</u>	1.162	1.267	1.590	1.688

Mean Reaction Time: dual-task Trials

Easy - 2 Targets	1.659	1.540	1.645	1.806
Difficult - 2 Targets	1.596	1.686	1.728	1.850
Easy - 4 Targets	1.631	1.560	1.659	1.838
Difficult - 4 Targets	1.643	1.743	1.799	1.750
<u>Mean: 2 Targets</u>	1.128	1.219	1.527	1.650
<u>Mean: 4 Targets</u>	1.263	1.352	1.430	1.590

DUAL-TASK TRIALS

Auditory Sternberg. Analyses of the data from the auditory Sternberg trials (see Table 3) indicated that speech intelligibility level affected both response accuracy,  $F(3,78) = 172.6$  and response latency,  $F(3,78) = 3.2$ . There was also a significant main effect of the number of target items in the accuracy data, with 82.4% correct responses in the two-alternative condition vs. 76.8% correct responses in the four-alternative condition,  $F(1,26) = 21.6$ . Finally, the Number of Target Items x Speech Intelligibility interaction was not statistically significant,  $F(3,78) = 2.5$ ,  $p = 0.06$ . Simple effects tests revealed that this interaction is attributable to the fact that there was no significant difference between the two- and four-alternative conditions at 80% intelligibility ( $p > 0.20$ ) while there was a significant (all  $p$ 's  $< 0.05$ , one-tailed) difference at the other three intelligibility levels.



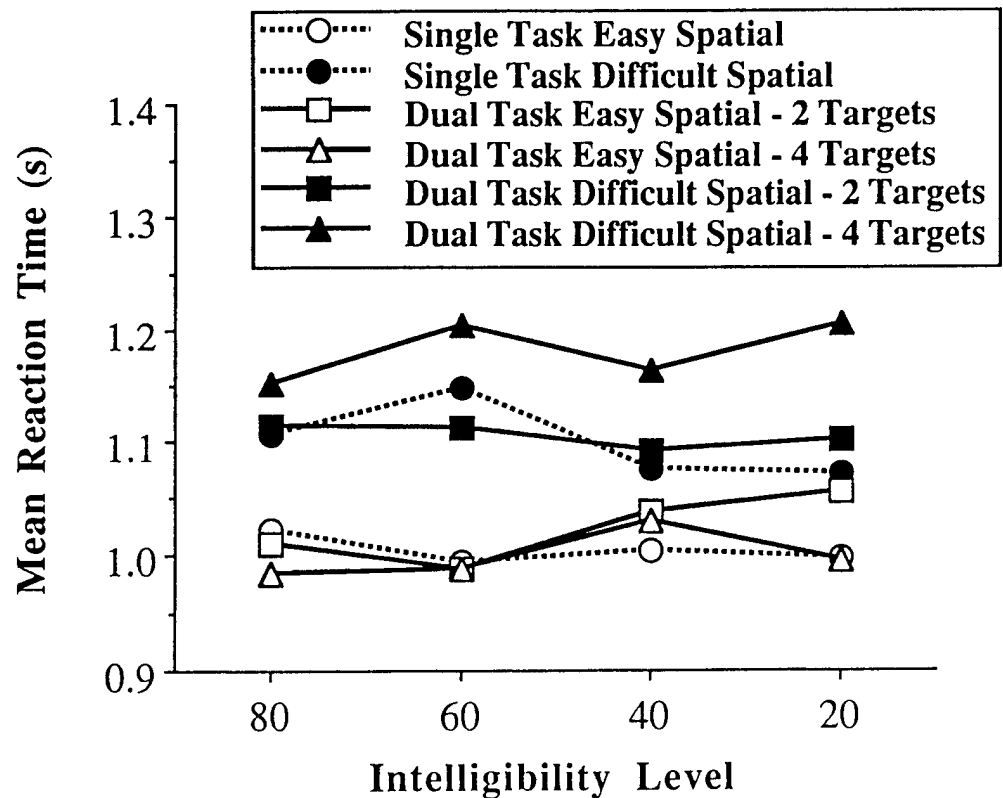
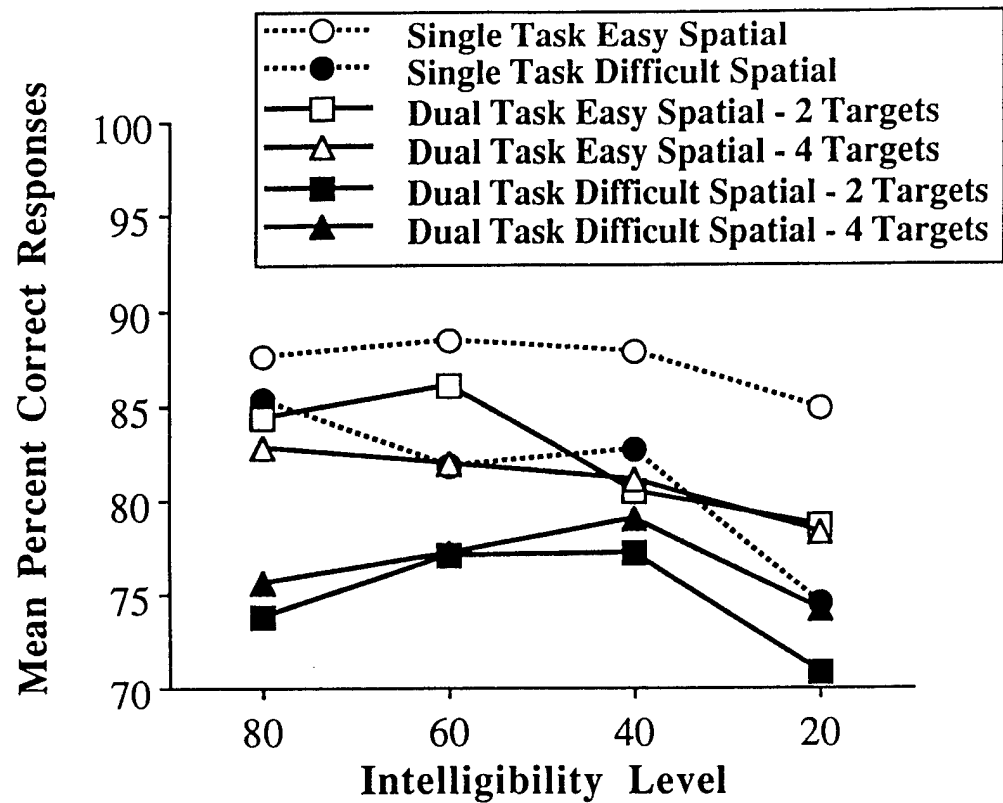


Figure 2. Experiment 2: Mean percent correct responses (upper panel) and mean RTs (lower panel) in the single and dual-task spatial processing trials for the easy and difficult spatial processing conditions.

**Spatial Processing.** For the percent correct measure, there was a significant difference between the easy and difficult conditions,  $F(1,26) = 10.6$ . More importantly, there was a significant main effect of speech intelligibility,  $F(3,78) = 3.0$ . Averaging performance across the easy and difficult spatial processing tasks, the mean percent correct responses for the 80%, 60%, 40%, and 20% speech intelligibility levels were 79%, 80%, 79%, 75%, respectively. Thus performance levels were consistent across the 80%, 60%, and 40% intelligibility levels, with accuracy decreasing only in the 20% speech intelligibility level condition.

For the RT measures the only notable finding was that subjects in the difficult task were slightly but not significantly ( $F(1,26) = 2.65$ ,  $p = 0.11$ ) slower than in the easy spatial processing task. There was no indication that speech intelligibility level affected RT ( $F < 1.0$ ).

## **Discussion**

The results from the dual-task trials indicate that at low levels of speech intelligibility performance in the spatial processing task was negatively affected. This impact was observed in the accuracy measure but not in the RT data. One aspect of the results of Experiment 2 warrants further attention: There was some indication of lower accuracy in the single-task spatial processing trials during the 20% speech intelligibility block of trials. This result is counterintuitive since, given that subjects were not performing the concurrent auditory task, one would not expect an effect of the level of speech intelligibility. There are three possible explanations for this finding. First, the result may represent sampling error. Second, recall that within each block of tasks performed at a given level of intelligibility there were two auditory tasks that preceded the single-task visual task trial (i.e., an MRT trial and a single-task auditory Sternberg trial). It is possible that when these two auditory tasks were performed at a very low level of intelligibility there was a carryover effect revealed in the single-task spatial processing trial. Third, it is possible that the decrease in accuracy in the single-task trials in the 20% intelligibility level condition reflects a speed-accuracy tradeoff since the decrease in accuracy was accompanied by a decrease in reaction time. The design of Experiment 2 precludes testing among these three views and, hence, deciding between these explanations will require additional research.

## **EXPERIMENT 3**

Experiment 3 was a replication and extension of the previous experiments with the sole difference being that the visual task used here was the CTS mathematical processing task. In this task subjects are presented with addition and subtraction problems, and their task is to determine whether the solution is greater than or less than five. This task taps the central resources of memory and decision-making and thus, according to multiple resource theory, is similar to the spatial processing task.

## Method

### SUBJECTS, DESIGN, AND APPARATUS

Twenty-eight students were recruited from the same source as the previous experiments. The experimental design was the same as in Experiments 1 and 2 with the exception that the visual task employed here was the CTS math processing task.

## Results and Discussion

### SINGLE-TASK TRIALS

Auditory Sternberg. As expected, performance levels (see Table 4) were affected by both the number of target items and by the level of speech intelligibility. There was a significant main effect of intelligibility level for both the accuracy,  $F(3,78) = 156.8$ , and RT measures,  $F(3,78) = 7.3$ . Also as expected, RTs were affected by the number of target items,  $F(1,26) = 10.2$ , with slower RTs in the four-item condition than the two-item condition.

Not surprisingly, the difficulty of the math processing did not affect performance in the single-task auditory Sternberg Task. The overall mean RTs for the easy and difficult math processing conditions were 1.335 and 1.378 secs, respectively. The corresponding data for the accuracy measure were 81% and 82%, respectively. There was no main effect of math processing difficulty for either the percent correct measure,  $F(1,26) = 0.8$ , or the RT measure  $F(1,26) = 0.3$ .

Math Processing. Presented in Figure 3 are the percent correct responses (upper panel) and mean reaction times (lower panel) for the easy and difficult math processing task at each level of speech intelligibility. As these data indicate, there were large differences in the performance levels for these two levels of difficulty of math processing. Furthermore, there was no effect of the level of speech intelligibility for the block of trials within which this single-task trial was run.

These observations were supported by the results of separate 2 (Math Processing Difficulty: Easy vs. Difficult)  $\times$  4 (Speech Intelligibility Levels: 80%, 60%, 40%, 20%) mixed-factor ANOVAs, one for each dependent variable. The two analyses produced identical patterns of results showing a significant main effect of math processing difficulty,  $F(1,26) = 109.8$ , and  $F(1,26) = 4.7$ , for the RT and accuracy measures, respectively. For neither dependent variable did the main effect of Speech Intelligibility or the Math Processing Task Difficulty  $\times$  Speech Intelligibility interaction approach significance ( $F$ 's  $< 1.23$ ).

### DUAL-TASK TRIALS

Auditory Sternberg. As indicated in Table 4, speech intelligibility level exerted a significant effect on response accuracy,  $F(3,78) = 230.4$ . The number of target items also had a significant effect on accuracy in the dual-task trials,  $F(1,26) = 26.6$ , although the Number of Alternatives  $\times$  Speech Intelligibility level interaction did not achieve significance,  $F(3,78) = 2.3$ ,  $p = 0.08$ . In Experiment 3 this marginal interaction was due to a floor effect on accuracy levels at

Table 4

Experiment 3: Mean Percent Correct Responses and Reaction Times (sec) for the Single and dual-task Auditory Sternberg Trials

<u>Difficulty Level of Visual Task/</u>	<u>Intelligibility Condition</u>			
<u>Number of Targets</u>	<u>80%</u>	<u>60%</u>	<u>40%</u>	<u>20%</u>
<u>Percent Correct: single-task Trials</u>				
Easy - 2 Targets	91	97	80	54
Difficult - 2 Targets	99	98	77	60
Easy - 4 Targets	96	92	78	57
Difficult - 4 Targets	96	91	83	56
<u>Mean: 2 Targets</u>	95	98	79	57
<u>Mean: 4 Targets</u>	96	91	81	56
<u>Percent Correct: dual-task Trials</u>				
Easy - 2 Targets	96	97	83	57
Difficult - 2 Targets	96	96	87	58
Easy - 4 Targets	90	88	70	54
Difficult - 4 Targets	93	83	74	51
<u>Mean: 2 Targets</u>	96	96	85	58
<u>Mean: 4 Targets</u>	91	85	72	52

(Table 4, continued)

<u>Mean Reaction Time: single-task Trials</u>				
Easy - 2 Targets	1.328	1.144	1.259	1.418
Difficult - 2 Targets	1.172	1.200	1.424	1.541
Easy - 4 Targets	1.295	1.384	1.392	1.461
Difficult - 4 Targets	1.282	1.399	1.395	1.612
<u>Mean: 2 Targets</u>	1.250	1.172	1.341	1.480
<u>Mean: 4 Targets</u>	1.288	1.391	1.393	1.536

---

<u>Mean Reaction Time: dual-task Trials</u>				
Easy - 2 Targets	1.464	1.443	1.439	1.454
Difficult - 2 Targets	1.624	1.731	1.754	1.799
Easy - 4 Targets	1.487	1.455	1.471	1.438
Difficult - 4 Targets	1.741	1.909	1.784	1.886
<u>Mean: 2 Targets</u>	1.544	1.587	1.600	1.626
<u>Mean: 4 Targets</u>	1.609	1.682	1.627	1.662

---

the 20% intelligibility level. The analysis of the RT data revealed a significant main effect for the difficulty level on the math processing task,  $F(1,26) = 12.9$ .

Math Processing. There are two important points to note regarding the RT data (see the bottom panel of Figure 3). First, as expected there was a large difference in the mean reaction times for the easy and difficult math processing task,  $F(1,26) = 76.2$ . Second, in the difficult math processing condition, reaction time increased as speech intelligibility levels decreased, whereas the reaction times in the easy math processing condition were approximately equal for the four speech intelligibility levels,  $F(3,78) = 2.9$ .

Turning to the accuracy data (see Figure 3, upper panel), it is clear that the level of speech intelligibility affected accuracy in the math processing task for both the easy and difficult math processing conditions. There were significant main effects for intelligibility level,  $F(3,78) = 76.4$ , and number of target items on the auditory Sternberg Task,  $F(1,26) = 270.0$ , as well as a significant number of target items x intelligibility level interaction,  $F(3,78) = 54.2$ . The nature of this interaction is illustrated in Figure 3. For the dual-task conditions there was only a small

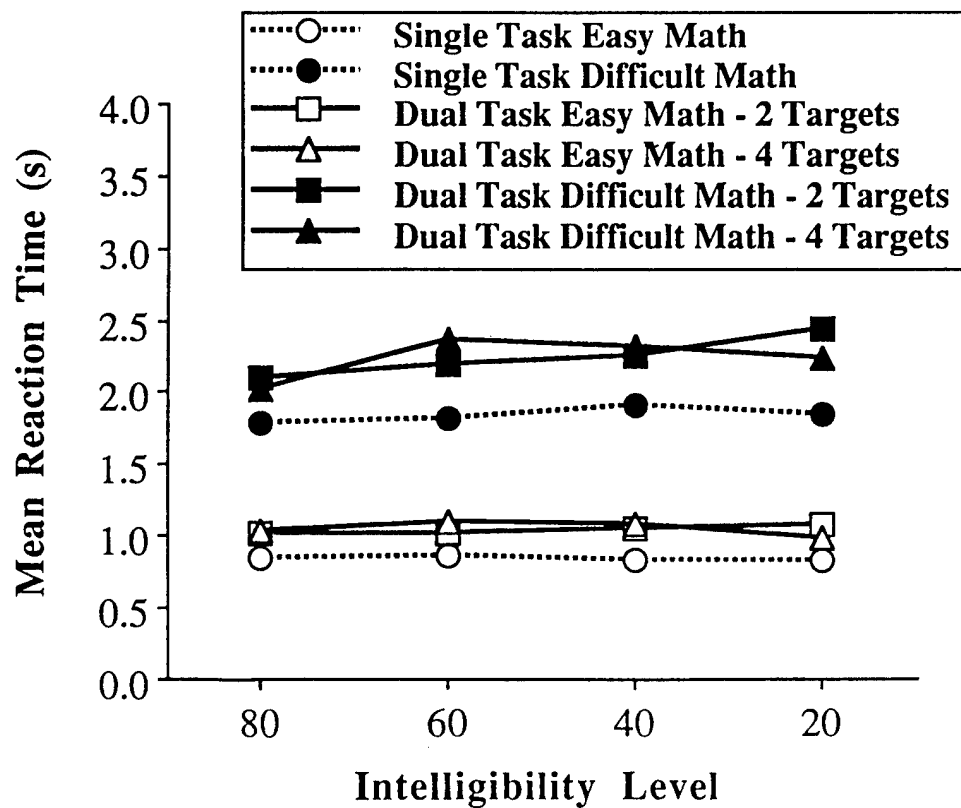
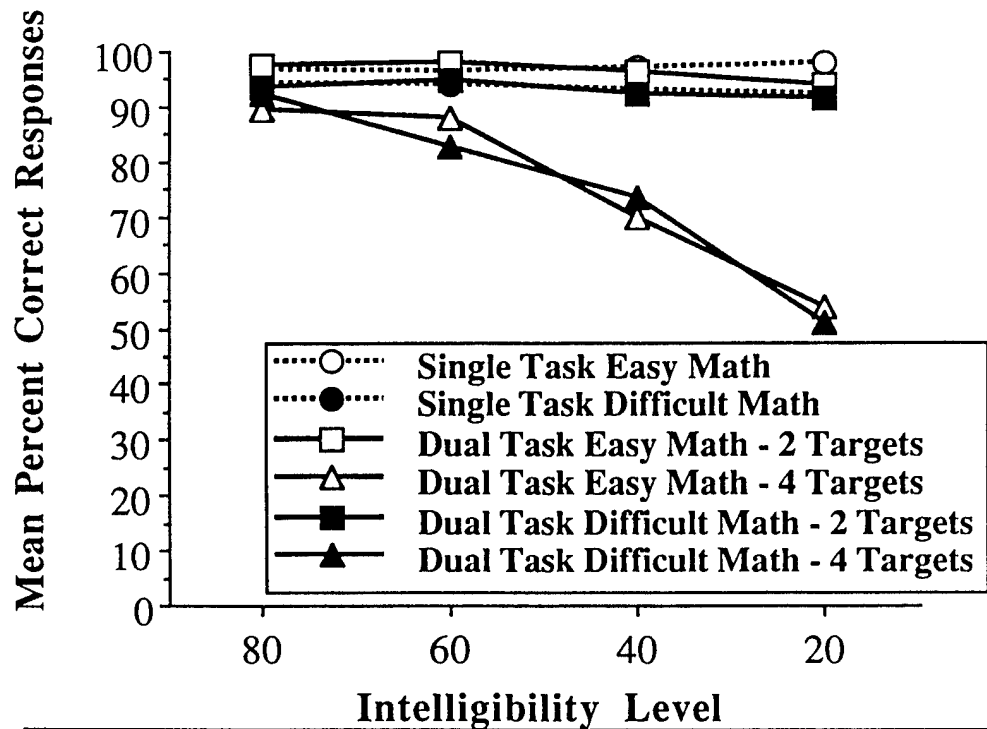


Figure 3. Experiment 3: Mean RT in the math processing task dual-task trials for the 2 and 4 target item conditions on the auditory Sternberg Task.

decrease in subjects' accuracy on the math processing task as the level of speech intelligibility decreased in the two-item condition, whereas there was a large decrease in performance in the four-item condition. Simple effects tests revealed a marginal effect of intelligibility level for the two-item condition,  $F(3,78) = 2.2$ ,  $p = 0.09$  and a significant effect in the four-item condition,  $F(3,78) = 84.8$ .

## **EXPERIMENT 4**

The results from Experiments 1-3 suggest that changes in speech intelligibility are most likely to impact performance in tasks that require central processes such as working memory or decision-making and are unlikely to impact performance in perceptual/motor tasks. There is, however, a major confound concerning the types of responses required in these tasks. In the unstable tracking task, subjects make continuous manual responses (turning a rotary knob). The auditory Sternberg Task also involved a button-press response and, based on Wickens' (1980, 1984) model, it is possible that the similarity of the responses required for the visual and auditory tasks is what produced the selective interference obtained across the experiments. That is, in those experiments in which the responses were similar in the visual and auditory tasks, there was a selective interference effect with the level of performance on the visual task being determined by the level of speech intelligibility.

Experiment 4 was designed to extend the first three experiments and to test whether the patterns of selective interference obtained in the initial experiments were due to (a) differences in the stages of information processing involved in the various tasks or (b) the confounding of response type across the experiments. Towards this end we used a task that, according to Shingledecker et al. (1982) and Schlegel and Shingledecker (1985), primarily taps visual encoding and response execution. The visual task was a probability monitoring task that required subjects to monitor dials on a computer screen.

## **Method**

### **SUBJECTS, DESIGN, AND APPARATUS**

Twenty-nine subjects were recruited from the same source as in the previous experiments. The experimental design and procedures were the same as in Experiments 1-3 with the exceptions that (a) the visual task employed here was the CTS probability monitoring task and (b) there were three levels of difficulty used with the probability monitoring task. There were eight subjects tested in the easy and difficult probability monitoring conditions and 13 in the medium condition.

### **PROBABILITY MONITORING TASK**

In the CTS probability monitoring task, subjects are presented with one, three, or four displays on the computer screen. Each display consisted of a rectangle containing a row of six short vertical lines with a seventh line above the row of six and centered between the third and fourth lines of the row of six. This seventh line indicated the midpoint of the row of six lines. Finally, an arrow pointing upwards was located below the row of six lines. During the trial this

arrow moved (at 0.5 sec intervals) across the display. The arrow's movement was either in the nonsignal state, in which case the arrow moved in random movements across the entire display, or in the signal state, in which case the arrow stayed predominantly on either the left or the right side of the display. The subjects' task was to detect when the display changed from the nonsignal to the signal state. Each of the displays was labeled with a number (1-4) and subjects were instructed to press the response button corresponding to the display that they thought was in the signal state.

Task difficulty was manipulated in two ways. First, the number of displays in the easy, medium, and difficult conditions was one, three, or four, respectively. Second, in the easy (1-display) condition the signal state corresponded to 95% of the arrow movements being on one side of the display with 5% on the opposite side. In the medium and difficult conditions the corresponding percentages were 85%/15% and 75%/25%, respectively.

## **Results and Discussion**

### **SINGLE-TASK TRIALS**

Auditory Sternberg. The single-task auditory Sternberg recognition accuracy and response latency data are presented in Table 5. Analyses revealed a significant main effect of intelligibility level for both the accuracy,  $F(3,78) = 173.1$ , and RT measures,  $F(3,78) = 20.2$ . Also, both RT and response accuracy were affected by the number of target items,  $F(1,26) = 51.4$ , and  $F(1,26) = 12.1$  respectively.

Probability Monitoring. Analyses of the RTs and the number of correct responses for the easy, medium, and difficult probability monitoring task at each level of speech intelligibility produced the expected single-task results. There was no main effect of level of intelligibility nor a Difficulty x Level of intelligibility interaction. However, as task difficulty was increased subjects made marginally fewer correct responses,  $F(2,26) = 2.6$ ,  $p = 0.09$ , and their RTs were significantly longer,  $F(2,26) = 11.7$  (see Figure 4).

### **DUAL-TASK TRIALS**

Auditory Sternberg. Once again, speech intelligibility level exerted a significant effect on response accuracy,  $F(3,78) = 131.5$ , and RTs,  $F(3,78) = 4.7$ . The number of target items also had a significant effect on accuracy,  $F(1,26) = 23.3$ , and RTs,  $F(3,78) = 26.8$ . For the RT measure there was a significant effect of the difficulty level of the probability monitoring task,  $F(3,78) = 4.6$ , as well as a Number of Alternatives x Difficulty Level of the probability monitoring task interaction,  $F(3,78) = 4.1$ .

Probability Monitoring. The results of the analyses of the RTs and the number of correct responses for the easy, medium, and difficult probability monitoring task may be summarized quite simply. There was a significant main effect of difficulty level in the probability monitoring task for both the number of correct responses,  $F(2,26) = 7.2$ , and the RT measure,  $F(2,26) = 10.4$  (see Figure 5).



Table 5

Experiment 4: Mean Percent Correct Responses and Reaction Times (sec) for the Single and dual-task Auditory Sternberg Trials

<u>Difficulty Level of Visual Task/</u>	<u>Intelligibility Condition</u>			
<u>Number of Targets</u>	<u>80%</u>	<u>60%</u>	<u>40%</u>	<u>20%</u>
<u>Percent Correct: single-task Trials</u>				
Easy - 2 Targets	99	99	90	60
Medium - 2 Targets	98	98	89	62
Difficult - 2 Targets	97	95	86	57
Easy - 4 Targets	98	95	75	62
Medium - 4 Targets	92	88	84	60
Difficult - 4 Targets	94	85	85	63
<u>Mean: 2 Targets</u>	98	97	88	60
<u>Mean: 4 Targets</u>	95	89	81	62
<u>Percent Correct: dual-task Trials</u>				
Easy - 2 Targets	95	96	87	58
Medium - 2 Targets	95	95	87	64
Difficult - 2 Targets	97	93	86	64
Easy - 4 Targets	92	89	81	65
Medium - 4 Targets	89	91	78	55
Difficult - 4 Targets	88	80	84	51
<u>Mean: 2 Targets</u>	96	95	87	62
<u>Mean: 4 Targets</u>	90	87	81	57

(Table 5, continued)

<u>Mean Reaction Time: single-task Trials</u>				
Easy - 2 Targets	1.121	1.176	1.184	1.309
Medium - 2 Targets	1.127	1.070	1.237	1.444
Difficult - 2 Targets	1.165	1.158	1.324	1.408
Easy - 4 Targets	1.213	1.240	1.326	1.337
Medium - 4 Targets	1.284	1.241	1.394	1.457
Difficult - 4 Targets	1.312	1.376	1.448	1.570
<u>Mean: 2 Targets</u>	1.138	1.135	1.248	1.387
<u>Mean: 4 Targets</u>	1.168	1.156	1.296	1.396
<hr/>				
<u>Mean Reaction Time: dual-task Trials</u>				
Easy - 2 Targets	1.339	1.380	1.316	1.408
Medium - 2 Targets	1.377	1.333	1.438	1.548
Difficult - 2 Targets	1.422	1.558	1.497	1.637
Easy - 4 Targets	1.337	1.388	1.512	1.477
Medium - 4 Targets	1.403	1.391	1.476	1.600
Difficult - 4 Targets	1.568	1.715	1.662	1.802
<u>Mean: 2 Targets</u>	1.379	1.424	1.417	1.531
<u>Mean: 4 Targets</u>	1.436	1.480	1.550	1.626
<hr/>				

The results of Experiment 4 indicate that it is the similarity of mental resources that determines the presence/absence of a cross-modal task interference effect. The probability monitoring task required a button-press response--the same as the response required by the auditory Sternberg Task--yet there was no evidence that the level of speech intelligibility affected performance in the probability monitoring task. In contrast, the spatial processing and math processing tasks both showed effects of the level of speech intelligibility. Thus what differs across tasks that either were or were not affected by the level of speech intelligibility is not the nature of the response but rather the nature of the mental resources required to perform the visual task.

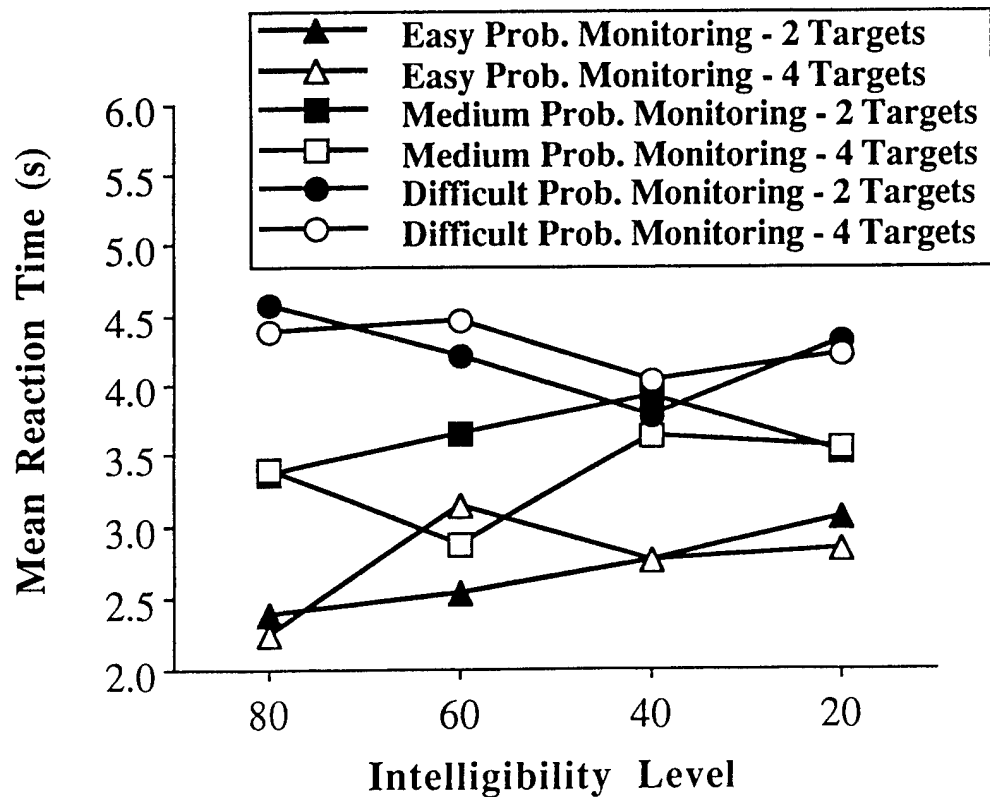
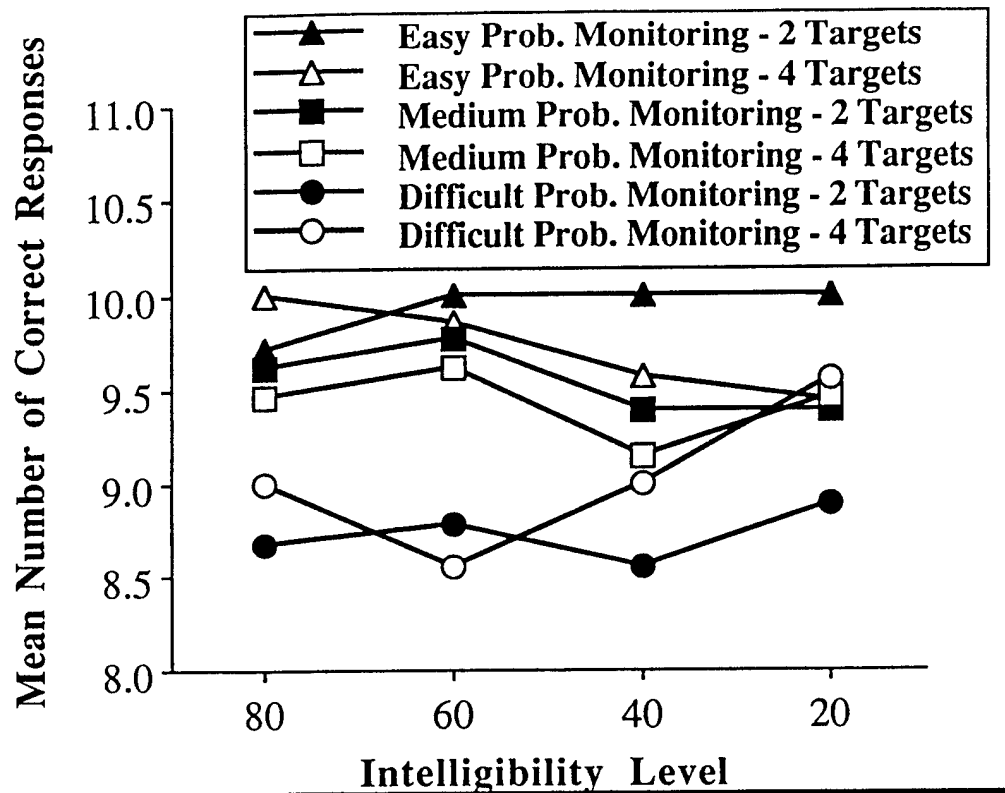


Figure 4. Experiment 4: Mean number of correct responses (upper panel) and RT (lower panel) in the probability monitoring task single-task trials for the easy, medium, and difficult probability monitoring conditions.

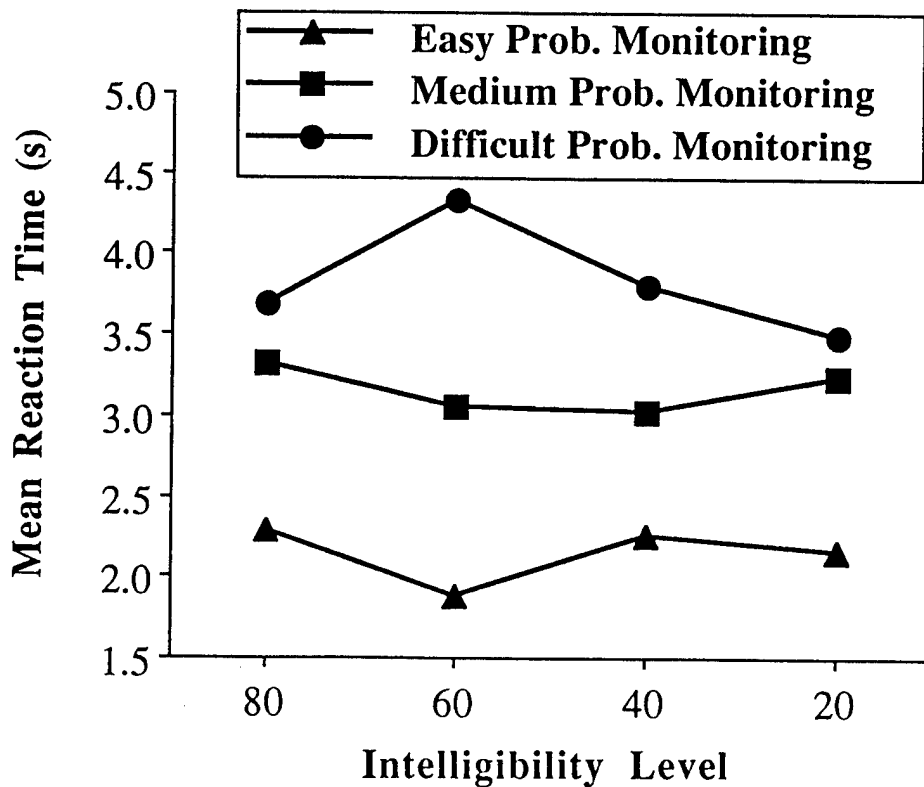
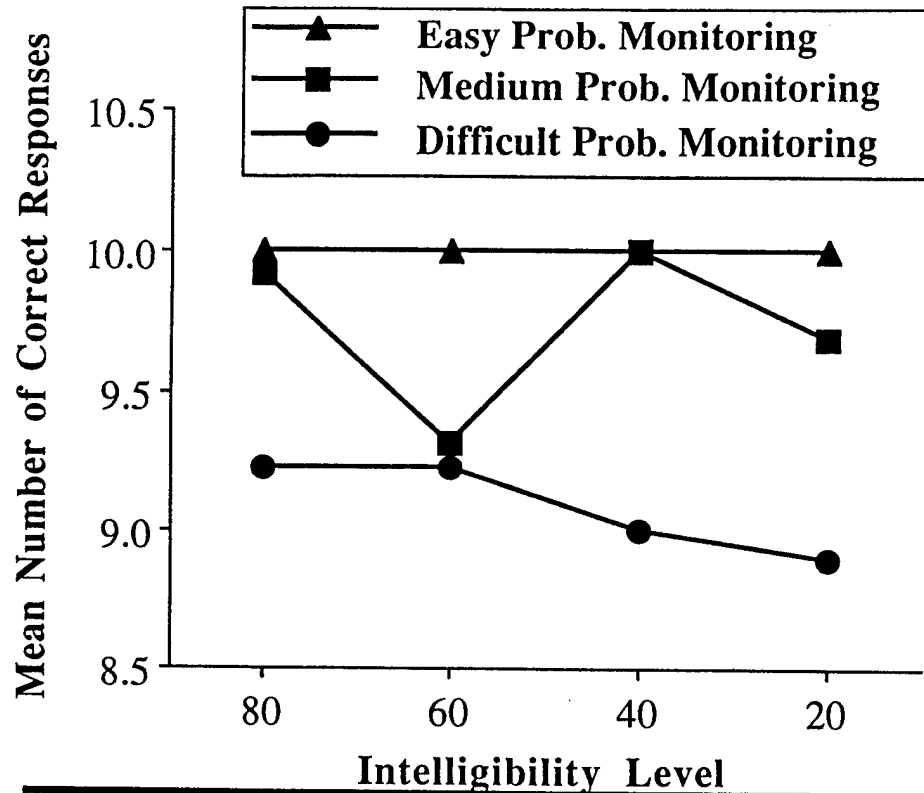


Figure 5. Experiment 4: Mean number of correct responses (upper panel) and RT (lower panel) in the probability monitoring task dual-task trials for the easy, medium, and difficult probability monitoring conditions.

## GENERAL DISCUSSION

The experiments reported here were designed to examine the extent to which changes in speech intelligibility levels would affect performance in concurrent visual tasks. Wickens' (1980, 1984) multiple resource framework was used to derive predictions concerning the patterns of selective interference that would be obtained with four different visual tasks. The results obtained here provide answers to several questions raised in the Introduction. First, the effects of changes in the intelligibility level for the auditory task did not produce a generalized interference effect, but rather the interference was restricted to those visual tasks that tapped the central processes of memory and decision-making. Although there were decrements in performance levels in the visual tasks in all of the dual-task conditions (relative to the single-task conditions), speech intelligibility levels affected performance levels only in the spatial processing and math processing tasks (Experiments 2 and 3). When the visual tasks tapped primarily perceptual and response resources (i.e., the tracking and probability monitoring tasks used in Experiments 1 and 4), there was no effect of speech intelligibility.

Taken together, these results are consistent with Wickens' (1980, 1984) multiple resource framework, and as such they provide support for that framework as a useful conceptual tool for examining cross-modal interference effects. The overall pattern of results from these experiments is inconsistent with single-capacity models (e.g., Broadbent, 1958) which postulate that attention/capacity can be allocated to any task(s). Exactly the same auditory task was used in Experiments 1-4 and, hence, the amount of "spare" capacity available to perform the visual tasks should have been the same in each experiment. In each experiment there was a decrease in performance levels from the single- to dual-task conditions, suggesting that the overall task demands in the dual-task conditions exceeded the available capacity. According to single-capacity models, then, there should have been an effect of speech intelligibility level in each of the four experiments. However, there was an effect of speech intelligibility level on performance in only two of the four experiments. In contrast to the single-capacity models, multiple resource theory exactly supports the results obtained here.

There are several other points to note with regard to the present research. First, based on the data presented here, it would appear that real-world activities that rely primarily on encoding and response resources would not be affected by the level of speech intelligibility. In contrast, we predict that activities that require memory and decision-making would be affected by the level of intelligibility.

A second point to note is that the impact of degraded speech intelligibility is greatest at intelligibility levels less than 40%. This indicates that future research could profit by concentrating more closely on performance levels obtained with intelligibility levels of less than 40%.

Third, the present results suggest that it may be possible to develop a battery of tasks that will allow researchers to identify individuals who are likely to be severely affected by lower intelligibility levels. Previous research has shown that performance in laboratory tasks can be effectively used to predict performance levels in real-world tasks such as flying aircraft (e.g.,

Gopher and Kahneman, 1971) or driving automobiles (e.g., Avolio, Kroeck, and Panek, 1985; Kahneman, Ben-Ishai, and Lotan, 1973). It would be of considerable use if an auditory test battery could be developed that would allow us to identify individuals whose performance is likely to suffer when intelligibility levels fall.

Finally, the results of the present experiments illustrate that the conceptual approach taken in this research has considerable applicability. It is important, however, that additional investigations be conducted to determine the generalizability of these results. Future research could vary, for example, the types of visual and/or auditory tasks employed in a dual-task procedure. It would also be of great importance to examine performance levels in more complex task environments in which operators are required to perform several tasks at the same time. Such a research approach would help to demonstrate the applicability of the present results to real-world analog tasks.

## REFERENCES

- Allport, D.A., Antonis, B., and Reynolds, P. "On the division of attention: A disproof of the single channel hypothesis," *Quarterly J. Exper. Psych.*, 24, 225-235 (1972).
- Avolio, B.J., Kroeck, K.G., and Panek, P.E. "Individual differences in information-processing ability as a predictor of motor vehicle accidents," *Human Factors*, 27, 577-587 (1985).
- Broadbent, D. Perception and communication. Oxford: Permagon (1958).
- Brooks, L.R. "Spatial and verbal components of the act of recall," *Canadian J. Psych.*, 22, 349-368 (1968).
- Goldinger, S.D., Pisoni, D.B., and Logan, J.S. "On the nature of talker variability effects on recall of spoken word lists," *J. Exper. Psych.: Learning, Memory & Cognition*, 17, 152-162 (1991).
- Gopher, D., and Kahneman, D. "Individual differences in attention and the prediction of flight criteria," *Perceptual and Motor Skills*, 33, 1335-1342 (1971).
- House, A.S., Williams, C.E., Hecker, M.H.L., and Kryter, K.D. "Articulation-testing methods: Consonantal differentiation with a closed-response set," *J. Acoust. Soc. Am.*, 37, 158-166 (1965).
- Kahneman, D., Ben-Ishai, R., and Lotan, M. "Relation of a test of attention to road accidents," *J. Applied Psych.*, 58, 113-115 (1973).
- Luce, P.A., Feustel, T.C., and Pisoni, D.B. "Capacity demands in short-term memory for synthetic and natural speech," *Human Factors*, 25, 17-32 (1983).

Martin, C.S., Mullennix, J.W., Pisoni, D.B., and Summers, W.V. "Effects of talker variability on recall of spoken word lists," *J. Exper. Psych.: Learning, Memory & Cognition*, 15, 676-684 (1993).

Navon, D., and Gopher, D. "On the economy of the human information processing system," *Psychological Review*, 86, 214-255 (1979).

Ogden, G.D., Levine, J.M., and Eisner, E.J. "Measurement of workload by secondary tasks," *Human Factors*, 21, 529-548 (1979).

Peters, L. and Garinther, L.J. "The effects of speech intelligibility on crew performance in an MIAI tank simulator," HEL Technical Memorandum (TM11-90). Aberdeen, MD: Aberdeen Proving Ground (1990).

Schlegel, R.E. and Shingledecker, C.A. "Training characteristics of the criterion task set workload assessment battery," *Proc. Human Factors Soc. 29th Annual Meet.*, Santa Monica, CA: Human Factors Society (1985).

Shaffer, L.H. "Multiple attention in continuous verbal tasks," in P.M.A. Rabbit and S. Dornic (Eds.), Attention and performance V, New York: Academic Press (1975).

Shingledecker, C.A., Crabtree, M.S., and Acton, W.H. "Standardized tests for the evaluation and classification of workload metrics," *Proc. 26th Annual Meet. Human Factors Soc.*, Santa Monica, CA: Human Factors Society (1982).

Sternberg, S. "The discovery of processing stages: Extension of Donder's method," in W.G. Koster (Ed.), Attention and Performance II, Amsterdam: North Holland (1969).

Whitaker, L.A., Peters, L.J., and Garinther, G. "Tank crew performance: Effects of speech intelligibility on target acquisition and subjective workload assessment," *Proc. Human Factors Soc. 33rd Annual Meet.*, Santa Monica CA: Human Factors Society (1989).

Whitaker, L.A., Peters, L.J., and Garinther, L. "Effect of communication degradation on military crew task performance," HEL Discussion Paper (DP 90-49), Aberdeen MD: Aberdeen Proving Ground (1990).

Wickens, C.D. "The structure of attentional resources," in P.S. Nickerson (Ed.), Attention and Performance VIII, Hillsdale NJ: Erlbaum (1980).

Wickens, C.D. "Processing resources in attention," in R. Parasuraman and R. Davies (Eds.), Varieties of Attention, New York: Academic Press (1984).

## **AUTHOR'S NOTES**

Portions of the research reported here were supported by the US Army Human Engineering Laboratory, Contract Number DAA15-89-K-0001, to David G. Payne. Requests for reprints should be sent to David G. Payne, Department of Psychology, SUNY-Binghamton, Binghamton, New York 13902-6000 (E-mail: [dpayne@bingvmb.bitnet](mailto:dpayne@bingvmb.bitnet)).



# **THE EFFECTS OF SPEECH INTELLIGIBILITY ON MILITARY PERFORMANCE**

**Georges R. Garinther**  
**US Army Research Laboratory**

**Leslie A. Whitaker**  
**Klein Associates**

**Leslie J. Peters**  
**97th General Hospital**  
**Frankfort, Germany**

## **BACKGROUND**

The ability of personnel to communicate clearly and to verbally coordinate crew operations is a critical factor for the successful completion of military missions. Degraded speech communication among crews of Army fighting vehicles during engagements may lead to misunderstandings, shooting the wrong targets, navigating to incorrect locations, mission failures, or even loss of life. Less than optimum communication may be the result of a number of different factors such as masking noise, hearing loss, a degraded hearing protector, a poor communication system, different dialects, etc.

Although it is readily apparent that good speech communication is necessary for the effective performance of armor crews, the quantitative relationship between speech intelligibility (SI) and mission success has never been established. It is also evident that complex tasks such as navigating through several checkpoints, identifying and shooting specified enemy targets, and providing verbal reports of friendly or enemy situations require highly effective communication. On the other hand, simpler tasks such as being in a stationary defensive position where the commander directs the gunner to identify and shoot a certain target can be accomplished at a lower speech intelligibility level (Garinther and Peters, 1990). Presently, communication systems in armored vehicles provide speech intelligibility levels of 55 to 75% when measured using the Modified Rhyme Test (MRT) (House, Williams, Hecker, and Kryter, 1963).

There is a need, therefore, to establish the SI requirements for the successful accomplishment of tasks at various levels of mission complexity for the following reasons:

1. Research has never been conducted to quantify the resulting performance at various levels of speech intelligibility. MIL-STD-1472D (1989) specifies the speech intelligibility levels necessary for systems that require communication. The MRT presently required for exceptionally high intelligibility is 97%, for normal intelligibility is 91%, and for minimally acceptable intelligibility is 75%. These requirements are educated estimates based upon the best information available.

2. Operations system analysts require parametric inputs if they are to conduct wargames to assess the effectiveness of our weapon systems against the enemy. At this time, if the analyst wishes to include the effect of communication upon mission outcome, the only data available is the speech intelligibility score. This is a meaningless parameter for wargaming. The analyst must be provided with performance parameters that address such factors as time to complete mission, percent of enemy vehicles killed, percent of own vehicles killed, number of checkpoints reached, etc.
3. Convincing performance data are necessary to indicate the need for providing improved communication in crew operated systems such as tanks. These data must be in a form that can provide those individuals responsible for procuring materiel the cost benefit alternatives relating to various levels of speech intelligibility.

## **METHOD**

### **General**

A series of five experiments was conducted between 1988 and 1993 to quantify the effects of communication upon armor crew performance. Typical parameters measured were:

- \* the time taken to conduct tasks such as identifying a target, conducting a mission, or relaying a status report;
- \* the percent completion of tasks such as enemy targets shot, checkpoints reached, or number of friendly vehicles hit by enemy fire;
- \* the number of errors made while performing tasks such as relaying a report or navigating to a specified checkpoint.

The performance level of these militarily relevant tasks was measured as a function of the crew's ability to communicate over the intercommunication system. We created communication-intensive scenarios which were conducted in tank simulators by the crews at four different levels of speech intelligibility: 25, 50, 75, and 100%. (The first experiment, gunnery only, also included scenarios conducted at a level of 7%.)

The experiments were conducted using professional tank crews with at least three years of experience in either the M1 Abrams or the M2 Bradley. These studies increased in complexity from scenarios that included only gunnery, to the addition of navigation, to the addition of subtasks such as transmitting situation reports, and finally to the conduct of force-on-force exercises.

### **Gunnery Scenarios**

These were relatively simple scenarios in which the tank commander had a script that required him to verbally instruct the gunner to shoot one of the four targets which were between one and three kilometers to the front of the tank. These four targets were a tank, a truck,

troops, and a helicopter. The gunner's task was to verbally confirm to the tank commander that he had identified the correct target and to shoot the target with the proper ammunition upon command from the tank commander.

### **Navigation, Reporting, and Gunnery Scenarios**

These were more complex scenarios in which a company commander, located in a command center outside of the tank, controlled the tank crew consisting of a tank commander, a gunner, and a driver. For each scenario, the company commander directed the tank commander to proceed through, and report in at, each of three checkpoints along a route of about three kilometers. Enemy vehicles were located at one of these checkpoints. The company commander instructed the tank commander to proceed along the route and to engage certain enemy vehicles; these instructions clearly stated that all other enemy vehicles were not to be shot. In addition, at one of the checkpoints the commanding officer requested the tank commander or the driver to provide a four-item report (the driver's report was requested through the tank commander).

### **Speech Intelligibility Test**

The MRT was used to measure speech intelligibility. The MRT consists of six lists of 50 monosyllabic English words. To establish the level of intelligibility, each crewmember read a list to all the other crewmembers. This was accomplished in round-robin fashion until everyone had read a list. The constant phrase, "would you mark \_\_\_\_\_ now" was used to present the specified monosyllabic word. The listener then selected the spoken word from among a closed set of six rhyming words printed on the answer sheet. The intelligibility score was the percent of words correctly chosen, adjusted for chance.

### **Subjects**

All crewmembers were screened for hearing in both ears to establish that they had thresholds not exceeding 25 dB at octave intervals from 250 Hz to 2000 Hz and 35 dB at 4000 Hz and 6000 Hz. Before the experiment the crews were trained until they consistently achieved a speech intelligibility score of at least 96% when using the MRT under ideal conditions. Typically the subjects underwent 2.5 hours of training. The crews were also trained to be equally proficient in the conduct of navigational exercises and scenarios which were similar to those used in the experiment.

### **Instrumentation**

Speech intelligibility was controlled by an electronic circuit that was used to set the desired level during each series of scenarios. This electronic circuit chopped the speech at a rate of 60 Hz, with the duty cycle (on-off time) of the speech being adjustable from 0 to 100%. Prior to these experiments, we had determined the duty cycle necessary to obtain each of the desired speech intelligibility levels. The circuit allowed the SI to be accurately set at the desired level for each test.

## Test Procedure

Immediately before a series of scenarios was conducted, the chopping circuit was adjusted to the nominal setting corresponding to the desired speech intelligibility level. Each subject then read a speech intelligibility list to the other crewmembers. If the scores fell within a preselected range of the nominal value, the scenarios began. If the scores were not acceptable, the chopping circuit was readjusted and the SI test repeated prior to the conduct of the scenarios.

The scenarios were presented in random order to the subjects at each level of speech intelligibility; speech intelligibility levels were presented to the subjects in counterbalanced order. Depending on the level of complexity, an individual scenario typically lasted about 10 to 25 minutes. After each series of scenarios, the intelligibility test was repeated, with the reported MRT score being the average of the pre- and post-tests.

## RESULTS

Typical results are shown from two of the five experiments that were conducted. Figures 1 and 2 are from very simple gunnery scenarios in which no navigation occurred and the crew was only required to shoot at a specified target. Figures 3 to 5 were from the more complex scenarios in which navigation, gunnery, and reporting occurred.

For the simple gunnery scenarios, Figure 1 shows the percent of time that the correct target was hit as a function of SI; these data show that percent of targets hit was fairly constant down to a level of 50%. Figure 2 shows the percent of time that the wrong target was hit as a function of SI; for these data the contrasts of 7% versus 25% ( $F=45.23$ ,  $p=0$ ) and 25% versus 50% ( $F=16.71$ ,  $p=0$ ) were significantly different.

For the more complex scenarios, Figure 3 shows the time required to complete the mission as a function of SI, Figure 4 shows the number of checkpoints reached as a function of SI, and Figure 5 shows the percent of missions successfully completed as a function of SI. Successful completion consisted of shooting the correct targets, transmitting most of the four-item report correctly, and reaching the final checkpoint.

## DISCUSSION

The results of the gunnery scenario indicate that missions requiring a limited dialog in which communications were often structured, e.g., "GUNNER-SABOT-TANK" (indicating that the gunner was to select a sabot round and shoot at a tank), produced a relatively high level of performance even at low speech intelligibility levels. Figures 1 and 2 indicate that the crews were able to perform with reasonable effectiveness at speech intelligibility levels as low as 50%; the percent of targets hit remained above 85%, and the percent of times that the wrong target was hit remained below 2%. As speech intelligibility was reduced to 25% and below, performance was dramatically reduced.

For the more complex navigation, gunnery, and reporting scenarios (Figures 3 to 5) in which communication was more interactive and in which sequential tasks occurred, performance

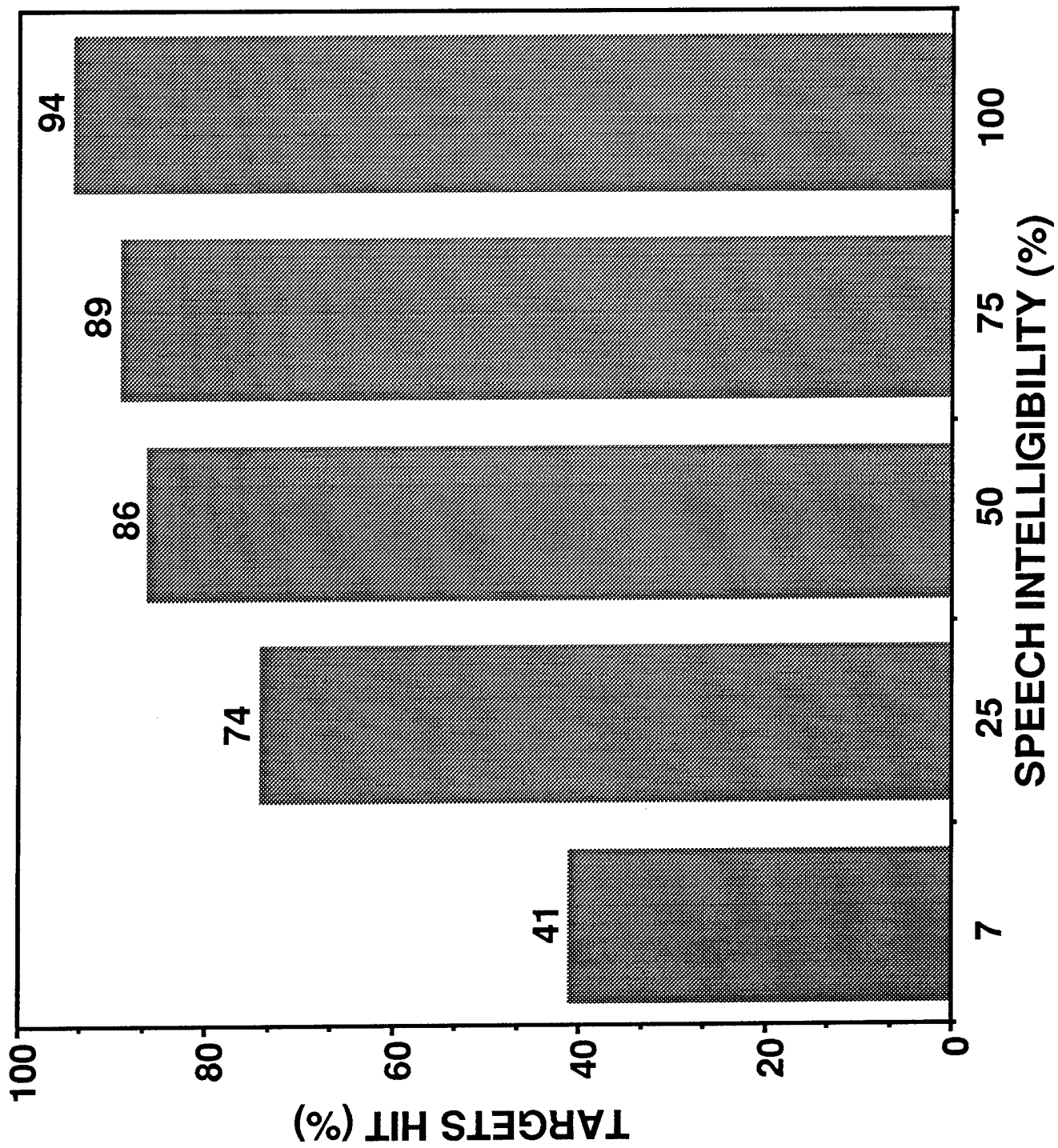


Figure 1. Percent of time the correct target was hit as a function of speech intelligibility.

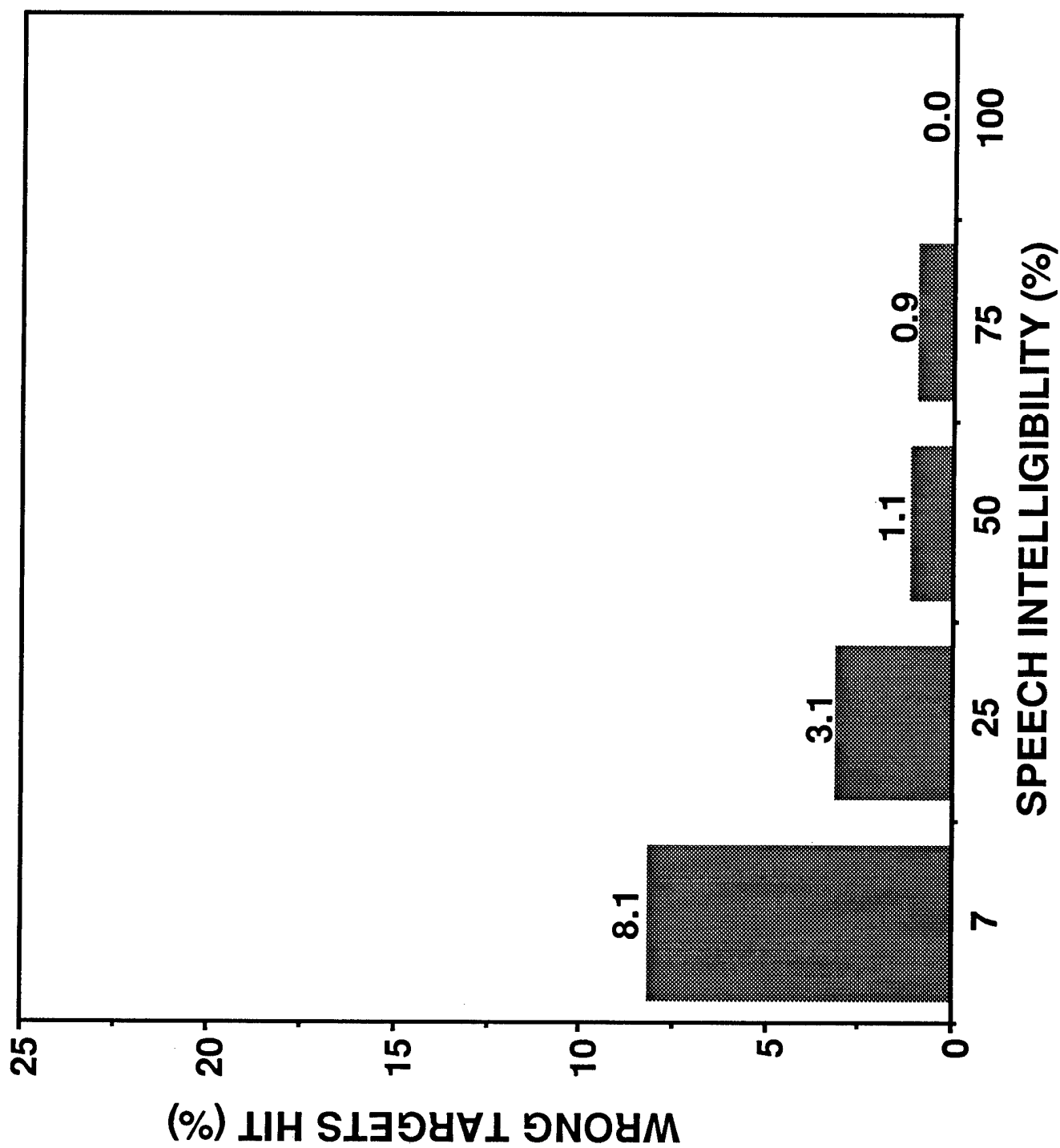


Figure 2. Percent of time the wrong target was hit as a function of speech intelligibility.

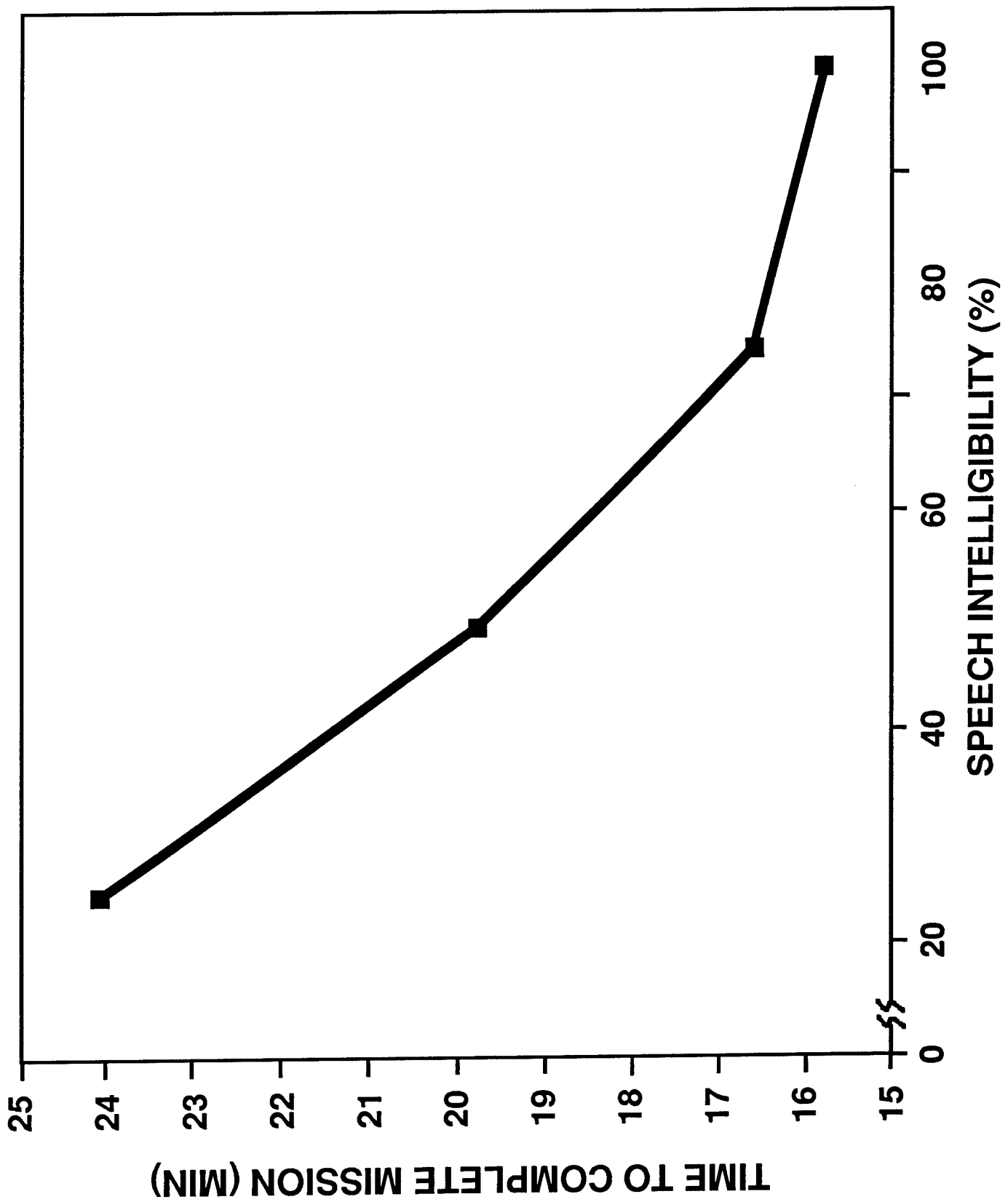


Figure 3. Time required to complete the mission as a function of speech intelligibility.

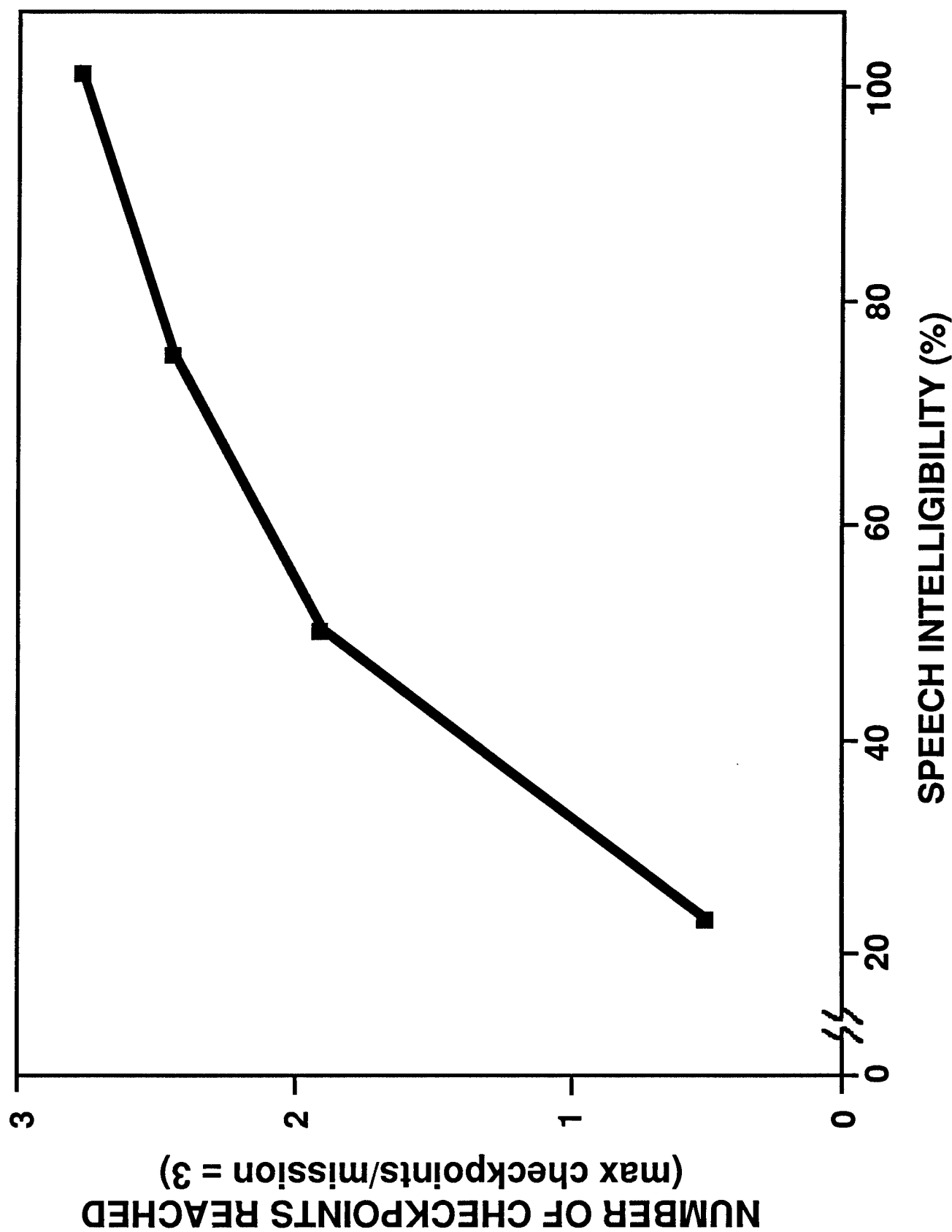


Figure 4. Number of checkpoints reached as a function of speech intelligibility.



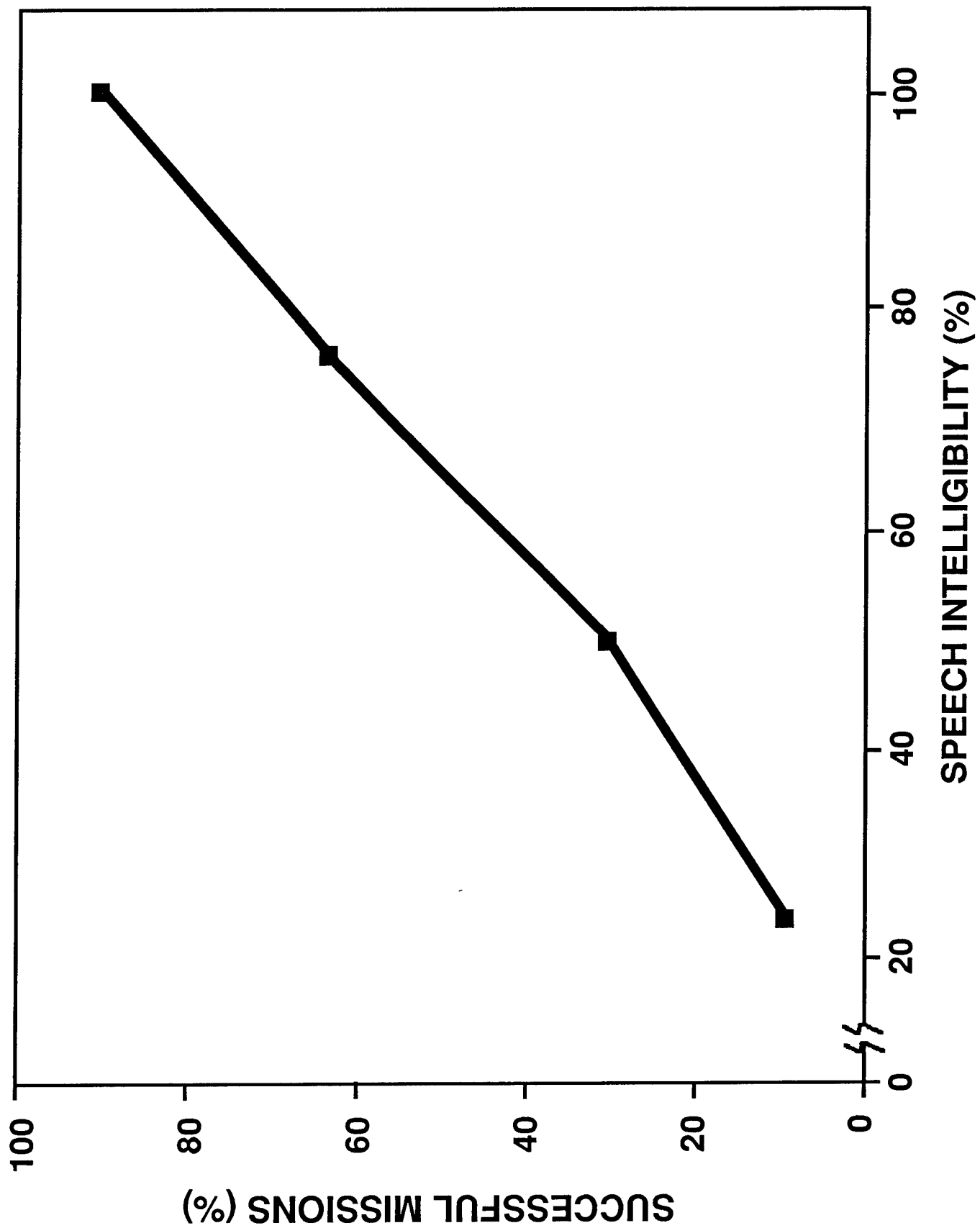


Figure 5. Percent of missions successfully completed as a function of speech intelligibility.

began to degrade at relatively high speech intelligibility levels. It is obvious that a task which is dependent upon the completion of a series of preceding tasks will have a lower probability of being accomplished. This reduced performance as a function of speech intelligibility is also dependent upon the complexity of the communication task (Whitaker, Peters, and Mitchell, 1992) and the structure of the communication (a command, a request for information, or a discussion).

Figure 3 indicates that the total time to complete a mission increased dramatically as speech intelligibility decreased. When speech intelligibility was reduced from 100% to 75%, the time increased by about 5%; from 75% to 25%, the time increased by about 45%. Figure 4 shows that the number of checkpoints reached decreased exponentially as speech intelligibility was reduced below 100%. Figure 5 shows that percent mission success decreased almost linearly with speech intelligibility.

A comparison of Figures 1 and 5 manifests most clearly the effect of increasing complexity from a simple scenario to one in which a series of dependent tasks must be accomplished using communication that is more complex and which requires a higher level of communication structure. Figure 1 shows an exponential decay with the knee of the curve at about 25%. Figure 5, however, shows an almost linear decay from a task performance of nearly 100% down to nearly 7%.

These figures show that complex scenarios requiring discussions among crewmembers, as, for example, when a company commander verbally describes the mission to the tank commander or when the tank commander describes a new route to the driver, necessitate high speech intelligibility if a high probability of mission success is expected.

An interesting added fact found in these studies was that about 10% to 15% of the crews were able to compensate, somewhat, for poor communication through various techniques. These techniques were only effective for situations that repeated themselves. This fact is mentioned here because it shows that troops will adapt with innovative procedures in combat situations, especially when their lives are in danger. For maximum combat effectiveness and crew survival, we should not rely on crew innovation, we must provide the crew with the best communication system available.

## CONCLUSIONS

A number of conclusions may be drawn from these studies:

- \* As scenario complexity increases, the performance of the crew begins to decrease at higher speech intelligibility levels. For simple gunnery scenarios, performance was reasonably effective at SI levels as low as 50% and was not dramatically reduced until the level approached 25%. For scenarios that included a series of sequential tasks and complex communication, crew performance decreased almost linearly as speech intelligibility was reduced.

\* The effects of speech intelligibility upon armor crew performance are significant and are measurable. Although it is self-evident that speech intelligibility will have an effect upon performance, no studies had been conducted that quantified these effects.

\* Communication systems that provide accurate speech intelligibility will increase combat performance. For complex communication requiring a series of sequential tasks leading to a mission goal, a specified percent improvement in speech intelligibility will provide an almost equal improvement in crew performance.

\* Both lives and materiel will be saved through improved communication, since more missions will be successfully completed.

\* The results of this study indicate that the Army's investment in improved communication systems will pay off in improved performance. Although armor personnel will adapt with innovative procedures to counter poor communications, maximum crew performance will be achieved by designing the best possible speech communication system into armored vehicles.

## REFERENCES

Department of Defense. Human engineering design criteria for military systems, equipment and facilities (MIL-STD-1472D). Washington DC (1989).

House, A., Williams, C., Hecker, M., and Kryter, K. Psychoacoustic speech tests: A modified rhyme test (Report ESD-TDR-63-403). Cambridge, MA: Bolt, Beranek, and Newman (1963).

Garinther, G., and Peters, L. Impact of communications on armor crew performance. Army Research, Development & Acquisition Bulletin, Jan-Feb 1990, 1-5 (1990).

Whitaker, L., Peters, L., and Mitchell, J. Measuring human performance as a function of speech communication using the Close Combat Test Bed facility. Proceedings of the 36th Human Factors Society Meeting, Atlanta GA, 12-16 Oct 1992, 237-241 (1992).

## **DEVELOPING A MESSAGE COMPLEXITY INDEX**

**Dr. Andrew Rose, Chief Scientist  
American Institutes for Research**

### **INTRODUCTION**

The American Institutes for Research (AIR) began an investigation of the relationship between the complexity of messages and performance under a contract with the Human Engineering Laboratory (HEL). The long-term objective of our program of research is to develop and test a model that relates message set complexity to performance. This program consists of three major phases. The major objectives of the first phase--the only phase completed--were as follows:

1. The first objective was to identify a set of message variables that we hypothesized would affect operational performance.
2. Once these variables were identified, the second objective was to develop an experimental paradigm for further exploration of these variables.
3. The third objective was to evaluate the variables in light of two primary considerations:
  - \* Can we define and measure each variable as a characteristic of operational messages? In other words, can the variable be applied to realistic (or all) messages?
  - \* Can we define and measure each variable in the context of messages to be used in the experimental paradigm?
4. The fourth and final objective of the first phase was to demonstrate the feasibility of a paradigm for investigating the relationship between the variables and performance. Could we collect meaningful performance measures that vary with changes in levels of message characteristics? Could the experimental paradigm support studies of one-way and two-way communications where the level of speech intelligibility is varied?

The major objective of the second phase of the research program would be to actually develop a quantitative model of message complexity through a series of laboratory studies. The basic approach would be to use the experimental paradigm developed in the first phase to collect performance data on a large number of messages differing with respect to the variables identified in Phase 1; these results would lead to scoring procedures for each variable, so as to allow us to generate performance predictions for other messages. In order to validate the complete prediction model, we would conduct a validation experiment in which messages were scored according to

the scoring procedures derived for each of the individual variable subsets. In the experiment, we would evaluate the goodness of fit of the model to the performance data, as well as the ability of the model to predict performance for new messages.

If successful, the result of the first two phases of the research program would be a laboratory-validated model. The third phase of the research program would be to integrate our research with other ongoing HEL-sponsored work and to validate the model against real-world operational performance. We hope to conduct experiments and observations of people performing actual communication tasks with measurable performance requirements. We would use the model to generate predictions, then evaluate the accuracy and reliability of those predictions. Whenever possible, the communication tasks would involve conditions of reduced (and, if feasible, controlled and measurable) degradations of speech intelligibility.

### **OBJECTIVE ONE: IDENTIFY MESSAGE VARIABLES**

We began the project by briefly reviewing experimental literature from various domains to identify variables affecting message complexity and performance. We grouped these variables on the basis of processes performed by the listener, within-message variables, and extra-linguistic or contextual factors.

As a result of our review and discussions between HEL and AIR, we decided to focus on within-message variables---basically syntactic and semantic features of messages---that have been shown, both empirically and theoretically, to affect performance. We concluded, however, that empirical support is lacking for several critical aspects of these variables. First, there has been little or no experimental work that establishes the relationships among these variables, especially when applied to auditory communication. In addition, there is little or no experimental work involving the interaction of any of these variables with levels of speech intelligibility, particularly given the way in which we will implement the degradations.

The current list of message variables is the following:

- \* Message Length
- \* Number of Ideas
- \* Word Frequency
- \* Redundancy
- \* Morphological Confusion
- \* Given-New vs. New-Given Order
- \* Expectancy
- \* Passive vs. Active
- \* Stative vs. Action Verb
- \* Personal vs. Impersonal
- \* Nominalization vs. Action Verb
- \* Levels of Subordination
- \* Type of Branching for Subordination

## **OBJECTIVE TWO: DEVELOP AN EXPERIMENTAL PARADIGM**

The second objective was to develop a paradigm to experimentally study message complexity. In addition to presenting messages that include levels of each of the variables in the current set, these experiments can be conducted while independently varying the level of speech intelligibility. We can also conduct these experiments within two communication structures, namely one-way and two-way configurations. A third configuration, multiple-path communications (e.g., discussions), was not addressed in this first round of experiments; however, the paradigm is sufficiently flexible to allow us to adapt it for the more complicated situation.

### **One-Way Communication**

In this paradigm, two subjects--the speaker and the respondent--are seated in two different rooms. They communicate by holding down a button while speaking into a microphone. The speech from the speaker to the respondent is filtered through the chopping circuit (under control of the experimenter). The settings for the chopping circuit are calibrated for the two subjects.

Both subjects face identical computer displays: an eight-by-eight grid, wherein each square is one of four colors (for example, blue, green, red, and yellow). A cursor appears in one of the squares; only the respondent controls the movement of the cursor by pressing the arrow keys on a keyboard. As an option in the program, the speaker's display can show where the respondent moves the cursor.

A typical trial would proceed as follows:

1. The speaker presses the microphone button.
2. A message appears on the speaker's screen. A typical message might be, "Move three squares north to the second yellow square."
3. The speaker reads the message into the microphone.
4. The respondent moves the cursor according to the directions heard in the message.
5. When the move is complete, the respondent presses a button and indicates readiness for the next message.
6. The speaker then presses the microphone button, and the next message is displayed.

A problem ends when the respondent finds the square that reveals the message, "End," or when the experimenter terminates it via the computer.

## **Hardware Configuration**

Two small offices, in close proximity to each other, were used for the experiments. We obtained two SSI Microfocus 386 System computers with ViewSonic four-color monitors and configured them for the experiments. The computers were set up and the programs were installed. The two computers were linked through serial communication ports.

## **The Display and Problem Presentation**

Each problem was designed around one of several predetermined paths through the grid, leading from a start square to an end square. Messages were scripted for each square that direct the respondent along the desired path. We have designed a flexible display that can present a grid with any number of rows and columns and with any of 16 colors in a square.

## **Responses and Dependent Variables**

The program was designed to record data on the accuracy of the speaker's message and the respondent's cursor movements. Within each trial the computer recorded:

- \* correct/incorrect cursor movements--whether the correct target square was achieved for each message
- \* the absolute error of the cursor position--the number of squares from the final cursor position to the correct target position

More globally, we also recorded the total time and the number of messages required to complete the problem.

## **OBJECTIVE THREE: DEFINE AND OPERATIONALIZE THE MESSAGE VARIABLES**

Given the list of message variables and the experimental paradigm described above, the next objective was to make the variables more concrete. For the variables listed above, we operationalized each as speaker messages in the experiment. For example, consider Message Length and Number of Ideas:

### ***Message Length***

**Definition:** The number of words in the message

**Levels =**

- 1) short = or < 8 words
- 2) medium = approximately 12 words
- 3) long = or > 16 words

**short:** Go left. Stop on the second blue box.

**medium:** Go left. Follow the most direct path. Stop on the second blue box.

long:            From your current position, go left. Follow the most direct path.  
                 Stop on the second blue box.

### ***Number of Ideas***

**Definition:**    The number of moves and supporting ideas in the overall message

Levels =        1) one move  
                 2) two moves  
                 3) three moves

one move:      Go left to the second blue box.

two moves:     Go up to the first red box. Go left to the second blue box.

three moves:   Go down to the first yellow box. Go left to the second green box. Go up  
                 to the first blue box.

Conceptually, all of these variables are independent: it should be possible to construct messages with all combinations of all of the levels of each variable. We have not attempted to examine all possible combinations. It is also true that many of these variables covary naturally in real messages; for example, more redundancy in a message will usually mean a longer message. Similarly, some combinations of variable levels are unlikely ever to occur operationally (e.g., short, redundant messages conveying three different ideas).

## **OBJECTIVE FOUR: COLLECT PRELIMINARY FEASIBILITY DATA**

The last objective was to collect data from experimental trials that would provide heuristic information about the variables, the experimental paradigm, and directions for work during the second and third phases of the research program. Below, we describe some basic experimental considerations.

### **Independent Variables**

The main independent variables were the content of the messages and the level of speech intelligibility. During the first phase of the research program, we used all of the message variables in constructing the scripts (except the last two mentioned above). We used four levels of intelligibility: 25%, 50%, 75%, and 100%, as have been used by other researchers in the HEL program.

### **Subjects**

We used eight AIR staff members (none of whom worked on this project) as subjects. Each of the speaker-respondent pairs was "calibrated" on the chopping circuit using the Modified Rhyme Test for four levels of intelligibility.



## **Experimental Design Considerations**

Since the primary purpose for conducting the experiments was a feasibility demonstration, we did not plan an extensive analysis of the performance data. We ran 25 problems per session, where each level of each message variable was included. Each session took approximately 30 minutes to complete.

## **General findings**

We collected information on several performance variables. These included:

- \* Message Transmission Time (C) is the time interval between the commander's initial button press (which reveals the message) and the release of the button at the end of the message.
- \* Response Time (R) is the time interval between the release of the commander's button at the end of the message transmission and the respondent's first cursor movement. Conceptually, this is the time necessary to process the message, decide upon the movement, and initiate the movement.
- \* Movement Time (R) is the time interval between the beginning and the end of the respondent's movement.
- \* Move-Mike Time (R) is the time interval between the end of the movement and when the respondent presses the button to begin transmitting the respondent message.
- \* Message Time (R) is the time interval between the respondent's initial button press (which reveals the message) and the release of the button at the end of the message.
- \* Time to Next Message (C) is the time interval between the end of the respondent's message and the commander's button press to reveal the next message.
- \* Percent of Targets Hit is the percentage of messages receiving correct responses.

One result that was apparent from a surface examination of the data was that, in terms of hitting the target, subjects performed excellently. Variation in speech intelligibility did not matter, even at a 75% reduction level. In future experiments, it will be important to systematically vary intelligibility, making sure to explore the boundary where effectiveness of communication begins to deteriorate.

Generally, there was a consistency among the various dependent measures for each variable. For example, for the Number of Ideas variable, results show that as the number of ideas in a message increases,

- \* Message Transmission Time increases,
- \* Response Time increases,
- \* Movement Time increases, and
- \* Percent of Targets Hit decreases.

This pattern (with the exception of the Percent of Targets Hit measure) occurred consistently for most of the variables.

### **Specific Variables**

With the caveats associated with small sample sizes and other limitations of this trial experiment, the initial trends indicated by an examination of the individual variables are intriguing. All of the variables have produced theoretically and practically interesting results. For example, Length of Message, while resulting in longer transmission time, seems to have allowed subjects to prepare to move more quickly; response times were shorter for longer messages. That this was not just a function of the number of words in the sentence is reflected in the Number of Ideas results: here, longer messages required substantially longer response time than shorter messages. Redundancy, which also involves increasingly lengthy messages but with no new information, resulted in slightly lengthened response times. If these results hold up with larger samples, we will be able to make some important inferences about processing of auditory information.

Also very encouraging are the preliminary results from other individual variables. Word Frequency, Passive vs. Active, Stative vs. Action Sentence, and Personal vs. Impersonal seem to have substantial effects on response time. As mentioned previously, there is very little information in the experimental literature that demonstrates these effects for orally-presented information. The results for Nominal vs. Action Verbs suggest that this might be an even more important variable than originally supposed; it seems as if this type of message combines the effects of word frequency and personal vs. impersonal messages. Likewise, the Given-New vs. New-Given Order seems to affect performance; subjects reported (and it was confirmed by other data) that presenting the target at the end of the message caused them to "lose" the specific required path. Again, it is premature to speculate on the reliability of all of these effects, particularly with respect to their direction and magnitude, when subjects will have to perform with more severe degradations of intelligibility.

# INDIVIDUAL DIFFERENCES IN SPEECH PERCEPTION BY EYE AND EAR

Charles S. Watson

Hearing and Communication Laboratory  
Department of Speech and Hearing Sciences  
Indiana University  
Bloomington, Indiana 47405

*Abstract.* Individual differences in speech processing abilities by both normal-hearing and by hearing-impaired listeners with similar audiograms do not appear to be predictable, or are only very weakly predictable, from differential abilities to detect spectral or temporal details of simple or complex nonspeech sounds. Even larger individual differences in speechreading (lipreading) abilities have been found to be difficult to predict on the basis of either intellectual or sensory abilities. It is suggested that at least a portion of the variance in speech processing by ear or by eye may be modality independent, possibly based on the ability to perceive linguistic "wholes" on the basis of linguistic fragments. Some new evidence is reported in support of this hypothesis, which may help to understand why hearing aids are so much more effectively used by some listeners than others. Such a modality-independent linguistic processing ability would also help to explain the remarkable range of speech-processing performance by patients with cochlear implants.

Under degraded listening conditions some normal-hearing people can understand speech considerably better than others. Similarly, some persons with impaired hearing succeed as hearing-aid users while others do not, despite similar hearing loss. The reasons for these differences remain unclear, despite considerable research and a variety of theoretical explanations. My colleagues and I became interested in this issue for several reasons, including (in no special order): (a) the large range of individual differences in listeners' performance that have been often observed in our work, and that of many others, on the perception of complex, nonspeech sounds; (b) the difficulty in demonstrating any clear associations between those individual differences in the abilities to discriminate differences in simple or complex nonspeech waveforms and differential abilities to process speech sounds; and (c) the practical possibility that whatever factors explain individual differences in speech processing might help to understand the poor success rates (approximately 50-60 per cent) for hearing aids, and the remarkably broad range of speech-recognition performance with cochlear implants.

Studies of listeners' abilities to process the details of complex nonspeech sounds have led many investigators to comment on the unexpectedly large range of performance among audiometrically normal listeners, compared to the range of thresholds for the discrimination or detection of simple stimuli (Grose and Hall, 1989; Green, 1988; Kidd, 1993; Neff et al., 1993; Watson, 1987; Wright and McFadden, 1990; among many others). But that observation has only rarely led to full-scale studies of

individual differences. For example, Green (1988, p. 94) deals with this problem by prescreening subjects for his profile experiments, testing only those who initially demonstrate discrimination of spectrally complex sound. The fact that some large percentage of candidates (sometimes as many as 50%) may fail such screening tests for either profile discrimination or for our tasks with tonal sequences is not viewed as a serious problem, since the research goal is an understanding of the detectability of spectral-temporal details of these classes of complex stimuli by those who can detect them. But the rejected subjects are normal-hearing college students who have no apparent difficulty understanding spoken English, and for that reason it might be asked whether the abilities to detect changes in profile stimuli or in tonal sequences are likely to be relevant to the understanding of speech. The point is not to criticize Green's interesting and theoretically provocative series of studies (or our own tonal-pattern experiments, for that matter). It is that they, like many other psychoacoustic investigations using small groups of selected subjects, provide an inadequate basis from which to estimate the distributions of auditory capabilities of the general population of listeners. To do so requires research that is explicitly designed to determine the range and nature of individual differences in large and representatively sampled groups of subjects.

There have been several efforts over the past half century to determine the primary variables, or "factors," in auditory discrimination abilities (reviewed in Johnson, Watson, and Jensen, 1987). Unfortunately, most of these studies suffer from weaknesses in the psychophysical methods employed (the older ones generally failing to control for response-bias effects), in the level of training, in the number or representativeness of the tested samples of subjects, or in the nature and range of tasks included. Nevertheless, studies employing factor-analytic techniques (or some approximation to them) and large numbers of subjects, often implicate two or three relatively independent sources of variance in auditory processing. Typical lists include: frequency analysis, duration-intensity resolution (possibly a tradeoff in the case of briefer sounds), and in some cases a component that is specific to complex sounds (for which auditory working memory seems a reasonable first-order explanatory construct).

After an initial study of individual differences using a 22-test battery (Johnson, Watson and Jensen, 1987) we designed a shorter battery suitable for use with large numbers of listeners. The Test of Basic Auditory Capabilities (TBAC) is a recorded series of auditory discrimination tests designed for use in large-group studies (Watson, et al., 1982). The eight tests included in the TBAC were chosen partly on the basis of the results obtained by Johnson et al. and in consideration of the results of earlier test-battery studies, but also because of the growing interest in measures of temporal processing. Three single-tone discrimination tasks are used to obtain thresholds for increments in the frequency, intensity, and duration of 1.0-KHz tones. The second three tasks use tonal sequences to measure the abilities to discriminate complex sounds on the basis of rhythm (series of six 20-msec pulses), temporal order (series of four pulses with middle two ordered either AB or BA) and by the presence or absence of a single component in a "word-length" (450-msec) nine-tone sequence. The final two tasks are a syllable-sequence analog to the four-tone sequence task, (/fa/-/ta/-/ka/-/pa), and a subset of the Resnick et al. (1975) nonsense-syllable identification test. Christopherson and Humes (1992) recently showed the mean test reliability (estimated by Cronbach's Alpha) for all TBAC subtests to be 0.79.

Table 1a shows the range of performance of 127 normal-hearing listeners on the TBAC tests. Percent correct scores have been converted to threshold, by fitting psychometric functions to the data for the discrimination tests (eight levels of difficulty are included on each test). Table 1b shows a factor

Test/Percentiles	10	25	50	75	90
Pitch [ $\Delta f$ , in Hz]	19.53	11.63	6.45	3.75	3.01
Intensity [ $\Delta I$ , in dB]	3.15	2.13	1.22	0.56	<0.5
Duration [ $\Delta T$ in msec]	64.7	46.3	30.4	23.4	19.2
Rhythm [ $\Delta T$ in msec]	20.3	13.7	9.7	6.99	5.67
Embedded Tone [ $\Delta T$ in msec]	77.1	57.8	39.8	33.0	22.3
Temporal Order for Tones [ $\Delta T$ in msec]	98.4	62.4	51.4	35.2	27.8
Temporal Order for Syllables [ $\Delta T$ in msec]	>250	217.3	163.5	125.0	85.9
Nonsense Syllable Identification P(C),	0.519	0.556	0.611	0.667	0.722

**Table 1a.** Population performance for 127 normal-hearing college students on the Test of Basic Auditory Capabilities (TBAC). Performance measures represent thresholds fitted to psychometric functions for subjects at the 10th, 25th, 50th, 75th and 90th percentiles of this population (Watson, et al., 1982).

Tests	Factor 1	Factor 2	Factor 3	Factor 4
Temporal Order (tones)	0.803	.....	.....	.....
Embedded Tones (9-tone sequences)	0.771	.....	.....	.....
Pitch (single tone)	0.659	.....	.....	.....
Temporal Order (syllables)	0.522	.....	0.478	-0.343
Duration (single tone)	.....	0.878	.....	.....
Intensity (single tone)	.....	0.731	.....	.....
Nonsense Syllable Identification	.....	.....	0.902	.....
Rhythm (pulse sequences)	.....	.....	.....	0.876

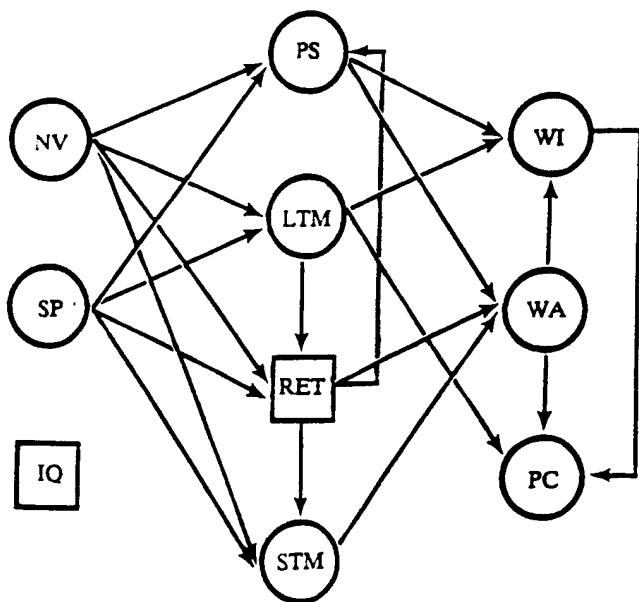
**Table 1b.** Sorted rotated factor loadings, based on arcsin-transformed percent correct scores on the eight subtests of the TBAC, for 127 normal-hearing college students. Loadings less than 0.25 are not shown (Watson, et al., 1982).

analysis of these data; factor loadings less than 0.25 are not shown. This analysis has been repeated for other large groups of subjects and in each case there appeared to be weak (or no) associations between the speech and nonspeech test scores (Espinoza-Varas and Watson, 1988). Espinoza-Varas et al. (1989) replicated this result with a group of 35 moderate-to-severe hearing-impaired listeners. In the latter study, the levels of the stimuli were elevated from the 75dB SPL previously used for normal listeners, to 15 to 30 dB SL for the impaired listeners. Performance means and standard deviations for the impaired listeners were not significantly different from those for the normals on the nonspeech tasks of the TBAC, while average performance on the speech tasks was slightly depressed.

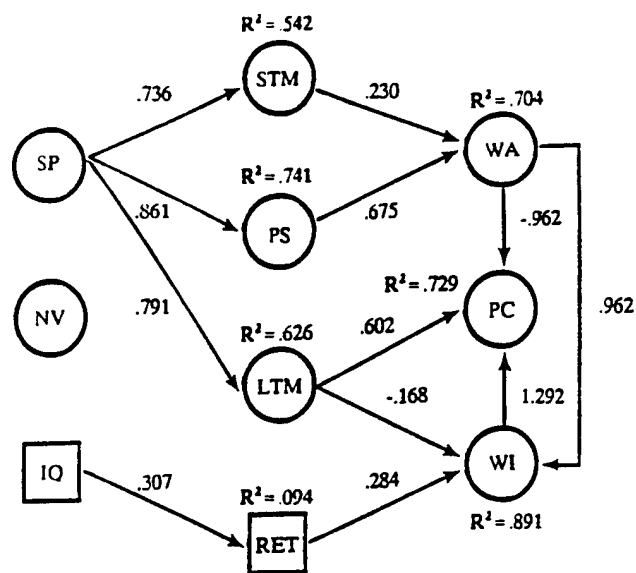
The TBAC has also been used to assess the auditory capabilities of subjects with reading or language disorders (B. Watson, 1992; B. Watson and Miller, 1993) and to examine the relation between intelligence and auditory discrimination performance (B. Watson, 1991). In general these studies show statistically significant but quite modest contributions of intelligence (e.g., correlations less than 0.3) to test performance. They also suggest that the hypothesized associations between language, learning or reading disabilities and auditory temporal processing is a very indirect one at best. The nature of this latter association was modeled by B. Watson and Miller (1992) using a structural-equation procedure, as illustrated in Figures 1 and 2. The best-fitting model suggests that temporal-processing abilities measured with nonspeech stimuli have a very minor association with language skills. Discrimination tasks with speech stimuli, on the other hand, are strongly related to linguistic tasks such as those used to measure phonological segmentation and short- and long-term verbal memory.

## **SENSORY AND COGNITIVE FACTORS RELATED TO INDIVIDUAL DIFFERENCES IN AUDITORY SPEECH PERCEPTION**

While there may be only very weak (or no) relations between performance on nonspeech discrimination tasks and speech perception, the ability to accurately recognize speech stimuli clearly does depend on auditory sensitivity. The most obvious and well-supported generalization about speech perception by ear is that "if you cannot hear it, you cannot understand it." Often you can hear some parts of the speech spectrum but not others, and the resulting perceptual errors are readily explained. These common-sense ideas are supported by a great deal of research that deals with the performance of *average* listeners as described, for example, in recent applications of the Articulation Index (e.g., Schum, Matthews and Lee, 1991; Pavlovic, 1984; Humes, et al., 1986), but also in the earliest systematic studies of speech perception (e.g., Miller and Nicely, 1955). Attempts to predict the speech perception performance of individual listeners have sometimes found reliable differences among persons whose auditory sensitivity and frequency resolving power are essentially identical (e.g., Plomp and Mimpen, 1979). Festen and Plomp (1983), however, argue that once the correlations with pure-tone sensitivity and with reduced frequency-resolving power are taken into account, there is only a small amount of variance in speech processing abilities among individual listeners. This view is largely based on samples in which there is a very broad range of auditory abilities, and it is possible that the differences in speech recognition accounted for by the degree of hearing loss may have simply "overwhelmed" the variance associated with nonauditory factors. Several other lines of research are consistent with this possibility.



NV Nonverbal temporal processing  
 SP Speech perception  
 IQ IQ  
 LTM Long-term memory  
 STM Short-term memory  
 PS Phoneme segmentation  
 RET Retrieval  
 PC Passage comprehension  
 WA Word Attack  
 WI Word Identification



NV Nonverbal Temporal Processing  
 SP Speech Perception  
 IQ IQ  
 LTM Long-term Memory  
 STM Short-term Memory  
 PS Phoneme Segmentation  
 RET Retrieval  
 PC Passage Comprehension  
 WA Word Attack  
 WI Word Identification

Figure 1. (left) Path diagram showing original form of Model 1, with all potential paths with the exception of those associated with IQ.

Figure 2 (right) shows the final model, preserving only the significant relations between auditory processing, phonological processing, and reading. (From Watson and Miller, 1993)

There is also evidence that cognitive abilities play a significant role in the processing of speech. In studies of hearing-impaired subjects, weak, but statistically significant relationships have been demonstrated between standardized measures of intelligence and various speech perception tasks. Stronger associations have been reported between various tasks measuring speed-of-information processing and speech perception (Knutson, 1988; Era et al., 1986; & van Rooij et al., 1989).

Van Rooij et al. (1989) found modest associations between measures of sensorimotor/perceptual speed and speech perception, once the effects of age were controlled, although hearing loss accounted for most of the variance in the speech perception scores of their subjects. More recently, van Rooij and Plomp (1992) reported a cognitive component of variance in speech perception for one group of elderly listeners but only sensitivity-related variance for a more severely affected group.

Similarly mixed reports exist regarding the abilities of hearing-impaired listeners to discriminate various spectral and temporal changes in nonspeech sounds, and of the associations between those abilities and speech recognition. For example, Dubno and Schaefer (1992) found reduced frequency selectivity for a sample of six impaired listeners, but no reduction in their abilities to identify noise-masked consonants--the latter presumably depending to some degree on frequency analysis. As noted earlier, using the TBAC test battery we have repeatedly failed to find significant correlations between the speech and nonspeech subtests, despite a wide range of performance and adequately reliable measures (Espinoza-Varas and Watson, 1988; Christopherson and Humes, 1992). One measure that has been reported to be degraded in the hearing impaired, and also to be correlated with speech reception, is the ability to detect or discriminate temporal gaps (e.g., Tyler, et al., 1982). This association has been questioned. Girandi-Perry et al. (1982) report animal data showing gap-threshold discrimination to be a strong function of the sensation level of the stimuli, which makes the interpretation of correlations between gap discrimination and speech perception more difficult, when listeners are tested with a broad range of sound levels. Moore and Glasberg (1988) found only slight differences between the gap discrimination abilities of normal and hearing-impaired listeners. Stelmachowicz et al. (1986) reported that the width of psychophysical tuning curves was strongly associated with speech perception in a mixed group of normal and hearing-impaired listeners, although most of the variance in both measures was for the hearing-impaired subjects. The only generalization that this literature seems to support is that if speech perception is strongly dependent on some specific general auditory ability (an ability that can be tested with nonspeech stimuli), that ability has yet to be discovered.

## SPEECHREADING

It is evident both from abundant research literature and from clinical practice that successful social communication by the hearing impaired depends on their abilities to speechread (lipread) as well as on their use of their residual hearing. Large individual differences in speechreading abilities, *even after training*, are well documented and in fact appear to be considerably greater than differences in the auditory speech processing abilities either of normal hearing subjects or of hearing-impaired subjects with similar audiograms.

After a century of correlational studies of psychological factors in speech-reading, investigators continue to ask the same question: How do good and poor speechreaders differ? That research has reached stronger agreements about what does *not* explain those differences than about what does. Many investigators have reported small but statistically significant correlations between psychological variables, such as general intelligence or linguistic knowledge, and speechreading, only to find that subsequent research failed to replicate their results (Gailey, 1987; also reviews by Jeffers and Barley, 1971, and by Berger, 1972).

There would appear to be at least several reasons that the correlational approach to the study of speech-reading has not led to more substantial and replicable findings. One is the variety of stimuli used in speech-reading studies. As Gailey (1987) and Demorest and Bernstein (1992) argue, it may be necessary to differentiate between types of speechreading, particularly speech-reading of more linguistically complex stimuli (e.g., sentences or stories) versus less complex stimuli (e.g., words or nonsense syllables). In a study with 40 normal-hearing college students, Gailey used five different



perception performance strictly from the audiogram. Some of those remaining differences seem likely to be related to cognitive or intellectual variables.

- (2) Psychoacoustic measures of acuity or resolving power with nonspeech stimuli do not correlate significantly with speech processing or correlate only weakly with it once the audiogram is taken into consideration, suggesting that a considerable amount of the systematic variance in speech-recognition abilities remains to be accounted for.
- (3) The cognitive skills shown to be significantly related to the broad range of individual differences in visual speechreading might also help to explain difference in the ability to recognize degraded speech waveforms presented to the ears.

## **EVIDENCE OF A MODALITY-INDEPENDENT SOURCE OF VARIANCE IN SPEECH PERCEPTION**

The lack of strong correlations between auditory discrimination measures and speech recognition led us to consider an additional source of variance that might contribute to individual differences in speech processing (Watson, 1991). A portion of the total speech-processing variance may not be modality-specific but rather involve a general ability to recognize linguistic "wholes" on the basis of linguistic fragments. Such an ability would help to explain the difficulty in finding strong associations between measures of speech processing and discrimination performance with nonspeech stimuli. We have recently found a simple way to demonstrate the existence of such a modality-independent source of variance. That approach has been to compare individual differences in visual speechreading to the same subjects' abilities to recognize speech in noise.

Many previous studies have measured speechreading performance and compared it to speechreading enhanced by auditory input, but we have thus far found only one that reported individual performance in speechreading and also in the ability to recognize speech strictly by ear (though we assume others must have done this). Dowell, Webb and Clark (1984) reported data for six cochlear-implant patients, and we calculated a correlation of 0.86 between their scores on a consonant identification test for look-alone vs. listen-alone. While this correlation is statistically significant, data from six implant subjects is scarcely a firm basis for a conclusion.

We therefore decided to look into the matter ourselves. In two MA thesis experiments by Mary Chamberlain and Weiguang Qiu in our lab at Indiana University, groups of normal-hearing college students were tested on two batteries of speech tests (this work is described in detail in an article currently under review). One series of tests involved video presentation of speech materials, including nonsense syllables, PB words, and CID sentences, and the other used similar materials presented acoustically, with speech-to-noise ratios low enough to insure that all subjects performed between chance and 100 percent correct, and thus could be ordered. Both studies, one with 40 subjects, the other with 50, yielded first-order canonical correlations between look-alone and listen-alone performance in the range of 0.45-0.55 ( $p < 0.01$ ). However, the amount of total testing was that which could be conducted in one two-hour session per subject, and the range of abilities was somewhat restricted because of the use of successful college students as the only subjects. For those reasons it is possible that the correlations between auditory and visual speech processing obtained with these

measures of speechreading which varied in linguistic complexity: syllables, words, sentences, familiar phrases, stories, and monologues. A factor analysis yielded two factors: one on which the simpler stimuli loaded (syllables, words), and another mainly related to the more linguistically complex measures (sentences, phrases, etc.). It is certainly plausible that different cognitive processes are required for the processing of simple and complex stimuli. General linguistic abilities and working memory (Baddeley, 1990) might be strongly related to performance in linguistically-complex speechreading tasks, while processing speed and perceptual closure might be more related to performance with less complex stimuli.

Gailey's studies, like so many others in the literature, have attempted to find the sources of individual differences in lipreading in cognitive or psychological strengths or weaknesses. The search for corresponding explanations of differential abilities to process speech through the auditory system has been much more concentrated on abilities assumed to reflect either peripheral auditory processing, perhaps reflecting psychoacousticians' rather modest interest in central or cognitive mechanisms. But there is no obvious reason that the cognitive processes shown to be significantly related to speechreading might not also account for individual differences in the auditory processing of degraded speech. For example, in Jeffers and Barley's older (1971) model of the major subskills underlying speechreading, a primary role was assigned to the accurate and rapid perception of the speech movements. As speech is encoded more rapidly, according to this model, other processing stages involving memory and comprehension may be made more efficient. There is no reason that these same arguments might not apply to processing speech presented to the ear.

## **VIBROTACTILE SPEECH PERCEPTION**

As in the cases of auditory discrimination abilities for speech and nonspeech stimuli, and for speechreading, there is a wide range of ability among users of tactile speech aids, even among users of the same device trained in optimized laboratory conditions. Bernstein et al. (1991) and Weisenberger (1991) have both raised the question of how this variability in performance might be explained. Even normal-hearing adult subjects trained with identical stimuli presented by the same talkers, for the same periods of time show a large range of performance (Weisenberger, et al., 1989; 1991). In a remarkable demonstration of the extremes in tactile abilities, Craig (1977) found two observers who were unexplainably superior to a large number of others tested in the recognition of text materials presented as vibrotactile patterns to the fingers (with an Opticon reader). Trained only briefly on the recognition of single letters, those two observers were able to read text at rates as high as 80 words per minute. Experienced blind users of the Opticon, even after years of practice, seldom achieved rates that high.

## **SOME PRELIMINARY CONCLUSIONS**

There is insufficient time today to review all of the literature on these matters. A very skeletal summary with which I believe at least some of the experts present in this audience will agree is:

- (1) While a very large portion of the variance in auditory speech processing can be associated with the audibility of the components of the speech waveform (e.g., with AI-based predictions), it is not possible to predict *individual differences* in speech

samples may underestimate that for the general population. The one aspect of the correlation that we could further analyze was its attenuation as a result of the level of test-retest reliabilities of the auditory and visual tests (using Cronbach's Alpha), however the reliabilities of the overall test battery were unfortunately high, meaning that the "true" correlation for this population of listeners is probably no greater than about 0.60. Accounting for 36 percent of the total variance is hardly a remarkable result, but again note that these were all normal hearing, successful college students and the range of auditory speech processing abilities is almost certainly smaller than that in the general population.

These results are consistent with the existence of a moderate-sized component of variance among normal subjects that is common to auditory and visual speech processing. The next question is, what other abilities might tap that same source of variance? Preliminary answers are most readily available from the lipreading literature, since those investigators have devoted such a considerable effort to the search for the linguistic, cognitive, or intellectual correlates of individual differences (Lyxell and Ronnberg, 1987).

As a first effort along those lines, we included, in Bill Qiu's study, an orthographic fragments test, as shown in Figure 3. Performance on thirty such sentences, scored either for total sentence correct, or for individual words correct, correlated about as well with both the composite visual speechreading score and the auditory composite score as those two forms of speech recognition correlated with each other (correlations of 0.4-0.5). It will also be interesting to see whether the cross-modality correlations are also found in the case of tactile speech processing.

This is clearly not a finished story. We think the cross-modality correlations help to explain (a) why we have such a difficult time accounting for individual differences in auditory speech processing in terms of strictly auditory abilities, and also (b) why it is that individual hearing-impaired listeners with the same audiograms, and apparently with the same information available through well-fitted hearing aids (or cochlear implants), sometimes differ so greatly in their abilities to recognize speech.

## **ACKNOWLEDGEMENTS**

The ideas presented here have evolved through discussions with too many colleagues to list, but I particularly appreciate many helpful comments from Gary Kidd, Diane Kewley-Port and Aimee Surprenant. Experimental work was conducted with the support of the National Institute for Deafness and other Communicative Disorders, and the Air Force Office of Scientific Research.

FISH CAN SWIM BUT  
 CAN'T SEE.  
 HE MISUNDERSTOOD  
 MY INSTRUCTIONS.  
 THE ALPINE, HAS  
 TWENTY-FOUR HOURS.

Figure 3. Samples of orthographic fragments task used by Qiu (1992).

## REFERENCES

- Baddeley, A. **Human Memory**. Boston: Allyn and Bacon (1990).
- Berger, K.W. *Speechreading: Principles and Methods*. MD: Baltimore, National Education Press (1972).
- Bernstein, L.E., and Demorest, M.E. Johns Hopkins lipreading corpus I-II: Disc 1, [Videodisc]. Baltimore: The Johns Hopkins University (1986).
- Christopherson, L.A., and Humes, L.E. Some psychometric properties of the Test of Basic Auditory Capabilities (TBAC). *J. Sp. and Hear. Res.*, 35, 929-935 (1992).
- Craig, J.C. Vibrotactile pattern perception: extraordinary observers. *Science*, 196, 450-452 (1977).
- Demorest, M.E., and Bernstein, L.E. Sources of variability in speechreading sentences. *J. Sp. and Hear. Res.*, 35, 876-891, (1992).
- Dowell, R.C., Webb, R.L., and Clark, G.M. Clinical results using a multiple-channel cochlear prosthesis. *Proceedings of the Second International Symposium: Cochlear Implants*, (1983, Paris, France). Also in *Acta Otolaryngologica*, Suppl. 411 (1984).
- Dubno, J.R., and Schaefer, A.B. Comparison of frequency selectivity and consonant recognition among hearing-impaired listeners. *J. Acoust. Soc. Am.*, 91, 2110-2121 (1992).

Era, P., Jukka, J., Qvarnberg, U., and Heikkinen, E. Pure-tone thresholds, speech understanding, and their correlates in samples of men of different ages. *Audiology*, 25, 338-352 (1986).

Espinoza-Varas, B., and Watson, C.S. Low commonality between tests of auditory discrimination and speech perception. *J. Acoust. Soc. Am.*, 84, S143 (1988).

Espinoza-Varas, B., Watson, C.S., and Patterson, S.E. Discrimination abilities of impaired listeners compared to the range of variation in performance of normal listeners. *J. Acoust. Soc. Am.*, 85, S24 (1989).

Festen, J.M., and Plomp, R. Relations between auditory functions in impaired hearing. *J. Acoust. Soc. Am.*, 73, 652-662 (1983).

Gailey, L. Psychological parameters of lipreading skill. In B. Dodd and R. Campbell (Eds.), *Hearing by Eye: The Psychology of Lipreading*. Erlbaum Associates Ltd. (1987).

Girandi-Perry, D., Salvi, R., and Henderson, D. Gap detection in hearing-impaired chinchillas. *J. Acoust. Soc. Am.*, 72, 187-1393 (1982).

Green, D.M. *Profile Analysis: Auditory Intensity Discrimination*, Oxford Univ. Press, New York (1988).

Grose, J.H., and Hall, J.W. "Comodulation masking release using SAM tonal complex maskers: Effects of modulation depth and signal position." *J. Acoust. Soc. Am.*, 85, 1276-1284 (1989).

Humes, L.E., Dirks, D.D., Bell, T.S., Ahlstrom, C., and Kincaid, G.E. Application of the Articulation Index and the Speech Transmission Index to the recognition of speech by normal and hearing-impaired listeners. *J. Sp. and Hear. Res.*, 29, 447-462, (1986).

Jeffers, J., and Barley, M. *Speechreading (lipreading)*. Springfield, IL: Charles C. Thomas (1971).

Johnson, D.M., Watson, C.S., and Jensen, J.K. Individual differences in auditory capabilities. *J. Acoust. Soc. Am.*, 81, 427-438 (1987).

Knutson, J.F. Psychological variables in the use of cochlear implants: predicting success and measuring change. *Cochlear Implants*. NIH Consensus Development Conference (1988).

Lyxell, B., and Ronnberg, J. Guessing and speech-reading. *British Journal of Audiology*, 21, 13-30 (1987).

Miller, G.A., and Nicely, P.E. An analysis of perceptual confusions among some English consonants. *J. Acoust. Soc. Am.*, 27, 338-352 (1955).

Moore, B.C.J., and Glasberg, B.R. Gap detection with sinusoids and noise in normal, impaired, and electrically stimulated ears. *J. Acoust. Soc. Am.*, 83, 1093-1101 (1988).

Neff, D.L., Dethlefs, T.M., and Jesteadt, W. Informational masking for multicomponent maskers with spectral gaps. *J. Acoust. Soc. Am.*, 94, 3112-3126 (1993).

Pavlovic, C. Use of the articulation index for assessing residual auditory function in listeners with sensorineural hearing impairment. *J. Acoust. Soc. Am.*, 75, 1253-1258 (1984).

Plomp, R., and Mimpen, A.M. Speech-reception threshold for sentences as a function of age and noise level. *J. Acoust. Soc. Am.*, 66, 1333-1342 (1979).

Resnick, S.B., Dubno, J.R., Hoffung, S., and Levitt, H. Phoneme errors on a nonsense syllable test. *J. Acoust. Soc. Am.*, 58, S144 (1975).

Ronnberg, J., Arlinger, S., Lyxell, B., and Kinnefors, C. Visual evoked potentials: relation to adult speechreading and cognitive functions. *J. Sp. and Hear. Res.*, 32, 725-735 (1989).

Schum, D.J., Matthews, L.J., and Lee, F.S. Actual and predicted word-recognition performance of elderly hearing-impaired listeners. *J. Sp. and Hear. Res.*, 34, 636-642 (1991).

Tyler, R., Summerfield, A.Q., Wood, E., and Fernandes, M. Psychoacoustic and phonetic temporal processing in normal and hearing impaired listeners. *J. Acoust. Soc. Am.*, 72, 740-752 (1982).

van Rooij, J.C.G.M., Plomp, R., and Orlebeke, J.F. Auditive and cognitive factors in speech perception by elderly listeners. I: Development of test battery. *J. Acoust. Soc. Am.*, 86, 1294-1307 (1989).

van Rooij, J.C.G.M., and Plomp, R. Auditive and cognitive factors in speech perception by elderly listeners. III. Additional data and final discussion. *J. Acoust. Soc. Am.*, 91, 1028-1033 (1992).

Watson, B.U. Some relations between intelligence and auditory discrimination. *J. Sp. and Hear. Res.*, 34, 621-627 (1991).

Watson, B.U. Auditory temporal acuity in normally achieving and learning-disabled college students. *J. Sp. and Hear. Res.*, 35, 148-156 (1992).

Watson, B.U., and Miller, T.K. Auditory perception, phonological processing, and reading ability/disability. *J. Sp. and Hear. Res.*, 36, 850-863 (1993).

Watson, C.S. Uncertainty, informational masking, and the capacity of immediate auditory memory. In W. Yost and C.S. Watson (Eds.), *Auditory Processing of Complex Sounds*, Hillsdale, NJ (1987).

Watson, C.S. Auditory perceptual learning and the cochlear implant. *Am. J. Otol.*, 12, Supplement, 73-79 (1991).

Watson, C.S., Johnson, D.M., Lehman, J.R., Kelly, W.J., and Jensen, J.K. An auditory discrimination test battery. *J. Acoust. Soc. Am.*, 78, Suppl. 1, 71, S73 (1982).

Weisenberger, J.M. Issues in evaluating wearable multichannel tactile aids. *J. Acoust. Soc. Am.*, 89, 1958 (1991).

Weisenberger, J.M., Broadstone, S.M., and Saunders, F.A. Evaluation of two multichannel tactile aids for the hearing-impaired. *J. Acoust. Soc. Am.*, 86, 1764-1775 (1989).

Weisenberger, J.M., Craig, J.C., and Abbott, G.D. Evaluation of a principle-components tactile aid of speech perception. *J. Acoust. Soc. Am.*, 90, 1944-1957 (1991).

Wright, B.A., and McFadden, D. Uncertainty about the correlation among temporal envelopes in two comodulation tasks. *J. Acoust. Soc. Am.*, 88, 1339-1350 (1990).

## SEQUENCE COMPARISON TECHNIQUES CAN BE USED TO STUDY SPEECH PERCEPTION

Lynne E. Bernstein

Center for Auditory and Speech Sciences  
Gallaudet University  
800 Florida Avenue, N.E.  
Washington DC 20002

This presentation covers three topics: 1) a brief general description of sequence comparison; 2) a description of the development of sequence comparator for phoneme-to-phoneme sentence alignment; and 3) a brief report on some results obtained with the comparator. Sequence comparison methods appear to have been discovered independently in several countries during the late 1960s and early 1970s (Kruskal, 1983). The goal of sequence comparison is to obtain an alignment between strings for which elements may have been deleted, inserted, or substituted (either as exact matches or replacements). To accomplish this goal, sequence comparison comprises two parts: the concept of distance between elements (costs) and algorithms to minimize total distance between strings. The current application is based on the description of sequence comparison in Sankoff and Kruskal (1983).

My interest in sequence comparison arose in the context of research on sensory aids for profoundly deaf people. The main goal for these devices (such as tactile aids and cochlear implants) is to improve speech communication. Usually this means enhancing a subject's ability to lipread (speechread). In developing and testing sensory aids, it is desirable, therefore, to employ evaluation measures applied to connected speech. It is also desirable to employ a testing procedure that is simple and imposes few constraints on subjects. Asking a subject what had just been said seemed such a straightforward, simple approach. Traditionally, when open set identification of this kind is employed, results are scored in terms of words or keywords correct. However, we had observed responses to the task of lipreading with or without a sensory aid that contained few or no words correct and yet appeared to be phonetically similar to the stimulus. It was hypothesized that much could be learned by studying the patterns of errors in responses, were it possible to obtain a systematic means of phonemically aligning the stimulus with the response.

The following stimulus-response sentence pair contains several common characteristics of errors from lipreading and problems that must be solved in generating alignments.

Stimulus: Proofread your final results.

Response: Blue fish are funny.

The initial consonants /b/ and /p/ are visually similar, hence the predictable substitution. The /u/ in proof and blue are similar although spelled differently. It appears that a word boundary has been misparsed, such that the final /f/ in proof is identified as the initial /f/ in funny. Other correct phonemes occur in the stimulus and the response in roughly the same order and location, such as /rf/ in your final



versus are funny, although the words do not match. The obtained sequence comparator (Bernstein et al., 1993a) solved the above alignment in the following manner:<sup>1</sup>

Stimulus: pru f#rid#yur#fAnL#r|z^lts

Response: blu#f-IS#a-r#f^n-#-i-----

## AN EXTREMELY BRIEF INTRODUCTION TO SEQUENCE COMPARISON

As mentioned above, sequence comparison has two main parts, metrics for measuring distance or similarity and algorithms for minimizing distance between sequences. The data submitted to the sequence comparator are:

Stimulus sequence:  $\mathbf{a} = \underline{a}_1 \dots \underline{a}_m$

Response sequence:  $\mathbf{b} = \underline{b}_1 \dots \underline{b}_n$

The sequence comparator also needs costs for inserting a phoneme in a response, costs for deleting a phoneme from a stimulus, and costs for substitutions (exact matches or replacements). Like the data submitted to the sequence comparator, the costs are determined prior to initiating the alignment process. As a step in the solution to the alignment problem, a stimulus-response matrix is constructed whose cells are the entries  $(\underline{a}_i, \underline{b}_j)$ . Each of the cells is processed according to a recurrence algorithm whose goal is to obtain a minimum distance between sequences and a phoneme-to-phoneme alignment of sequences. The basic recurrence equation is:

$$\underline{d}_{ij} = \min \begin{cases} \underline{d}_{i-1,j} + \text{deletion of } \underline{a}_i \\ \underline{d}_{i-1,j-1} + \text{substitution of } \underline{a}_i \text{ with } \underline{b}_j \\ \underline{d}_{i,j-1} + \text{insertion of } \underline{b}_j. \end{cases}$$

The recurrence equation implies that cells with lower subscripts will be processed before cells with higher subscripts. At each cell, the minimization is used to decide whether the cell should result in an insertion, a deletion, or a substitution. When the final cell,  $(\underline{a}_m, \underline{b}_n)$ , has been processed, the value  $\underline{d}_{mn}$  is the minimum distance between the sequences. A pointer equation corresponds to the recurrence equation:

$$\text{pointer}(i,j) = \begin{cases} i-1,j & \text{OR} \\ i-1,j-1 & \text{OR} \\ i,j-1 & \text{if term above is minimum.} \end{cases}$$

For every cell that is processed, one or more pointers is generated to be used for constructing the alignment. Each of the three expressions on the right of the pointer equation corresponds respectively to the three alignment options in the recurrence equation. If more than one value is minimum in the

<sup>1</sup> Note that the transcriptions in this paper are given in the DECTalk single-character notational system.

recurrence equation, more than one pointer will be generated and subsequently a corresponding alignment. That is, several different minimal alignments can be generated. Pointers are followed beginning with the final cell that is processed, ( $a_m, b_n$ ). Figure 1 shows a fragment of a possible costs matrix; a stimulus-response matrix with minimal distances shown in the lower right-hand corner of each cell; and pointers corresponding to the manner in which the phonemes are to be aligned. (See Sankoff and Kruskal, 1983, for other examples using the same notation.)

## DEVELOPMENT OF A PHONEME-TO-PHONEME SEQUENCE COMPARATOR

Since the aim of the work was to study lipread sentences, it was necessary to obtain a distance metric that applied to visual-phonetic similarity. The literature on lipreading contains numerous studies whose goal was to define a unit of visual-phonetic similarity known as the *viseme*. Visemes are visual equivalence classes of putatively noncontrasting phonemes (see Fisher, 1968; Owens and Blazek, 1985). For example, /b p m/ are considered a viseme, because they are visually ambiguous according to typical criteria. An initial comparator used the simple recurrence equation above with a viseme-based costs matrix. Table 1 shows how costs were assigned (Bernstein et al., 1993a).

Table 1. Costs of seven types of elementary alignments.

<u>Type of Elementary Alignment</u>	<u>Example</u>	<u>Cost</u>
Exact match	a, a	
Substitution within a viseme group	b, p, m	1
Substitution within consonants, but across visemes	b, g	2
Substitution within vowels, but across visemes	a, i	2
Substitution of consonants for vowels and vice versa	a, b	3
Insertion of a vowel or consonant in the response		1
Deletion of a vowel or consonant in the stimulus		1

Evaluation of the comparator made use of data from 139 normal-hearing and normal-vision young adults who lipread the 100 CID Everyday Sentences (Davis and Silverman, 1970) recorded on video laserdisc. Subjects lipread each sentence, and then typed at a computer terminal what the talker had said. 12,291 responses were obtained. Responses were checked for spelling errors and were corrected whenever errors were unambiguously due to spelling. Each response sentence was then transcribed using DECtalk, a text-to-speech synthesizer that produces a quasi-phonemic transcription as one stage in its transcription process. Transcription errors were corrected. Then the transcribed stimulus-response sentence pairs were submitted to the sequence comparator. Alignments and various

## Alignment Example

COST MATRIX:

	b	l	u	-
b	0	3	4	2
l		0	4	2
u			0	2

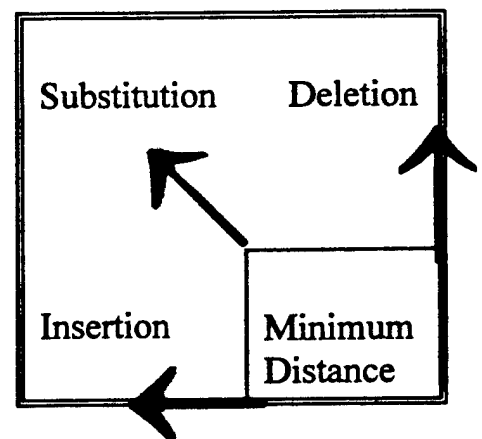
Stimulus: boo!

Response: blue

Response

Stimulus

		b	l	u
		0	2	2
		2	0	2
b		2	2	0
		2	4	2
u		4	2	2



Alignment:

Stimulus: b-u

Response: blu

Figure 1. Schematic representation of sequence comparison: a fragment of a hypothetical costs matrix, a stimulus-response matrix, and a diagram of the manner in which to interpret pointers.

measures were obtained, such as minimum distance, number of phonemes correct, number of phonemes deleted, and number of phonemes inserted.

Unfortunately, the combination of the simple recurrence equation and the viseme-based costs matrix resulted in inadequate constraint over the alignments. Numerous alternate equal-distance alignments were generated. Figure 2 shows four such alternate alignments for one sentence pair. Subsequently, several modifications described in detail in Bernstein et al. (1993a) were implemented. A more complex algorithm was implemented that charged an extra cost for initiating strings of insertions or deletions, thus reducing the use of insertions and deletions and the consequent fragmentation of response words. Still, however, an unacceptably high number of equal-cost alignments was obtained, although some improvement was achieved.

A costs matrix was then developed based on consonant confusions obtained in a nonsense syllable identification task. Multidimensional scaling was used to obtain Euclidean distances among consonants and among vowels. The new costs matrix provided additional resolution for reducing the number of equal distance alternate alignments. Unique alignments were obtained for 78% of stimulus-response pairs, and dual alignments (two alternate alignments) were obtained for 17% of pairs. A large proportion of the dual alignments involved a single elementary alignment, that is, one phoneme. The combination of the enhanced algorithm and the Euclidean distances was judged informally via inspection to be an adequate solution to the alignment problem.

## A VALIDATION EXPERIMENT

A problem for validating the sequence comparator was the absence of independently generated alignments with which the comparator's performance could be compared. It was not possible to validate the comparator against human judgments, since only the comparator could be expected to systematically and reliably obtain alignments. A different tack was taken, an evaluation of whether the comparator was sensitive to whether stimulus-response pairs were true or randomly assigned.<sup>2</sup> A main question was whether a large number of phoneme-to-phoneme alignments would be obtained regardless of whether the response was paired with its true stimulus.

The validation experiment used the same database of responses to CID Everyday Sentences as described above. One set of stimulus-response pairs were the true ones as collected in experimental sessions, and the other set were the randomly reassigned pairings. The results showed that exact phoneme matches were extremely rare in alignments of random pairs. 11,892 (96.7%) of randomly assigned sentence pairs resulted in five or fewer exact phoneme matches. 5,039 (41%) of true stimulus-response pairs resulted in six or more exact phoneme matches. Figure 3 shows the number of sentences as a function of number of phonemes correct for true and random pairs (Bernstein et al., 1993a).

<sup>2</sup> Note that the selection of response for each of the stimuli was random in the case of the random pairs. The alignment procedure operated identically on the random and true pairs.

## Alternate Equal-Distance Alignments

Stimulus:     Here's a nice quiet place to rest.

Response:     That's the way it goes.

### Alignment 1

Stimulus: hIrs-xnAskwA|tplestUrEst  
Response: D@tsDx----weItg-oz-----

### Alignment 2

Stimulus: hIrs-xnAskwA|tplestUrEst  
Response: D@tsDx----weIt----go--z-

### Alignment 3

Stimulus: hIrs-xnAskwA|tplestUrEst  
Response: D@tsDx----w-----eIt-goz-

### Alignment 4

Stimulus: hIrs-xnAskwA|tplestUrEst--  
Response: D@tsDx----weIt-----goz

Figure 2. Four alternate equal-distance alignments.

# Number of Sentences as a Function of Number of Phonemes Correct

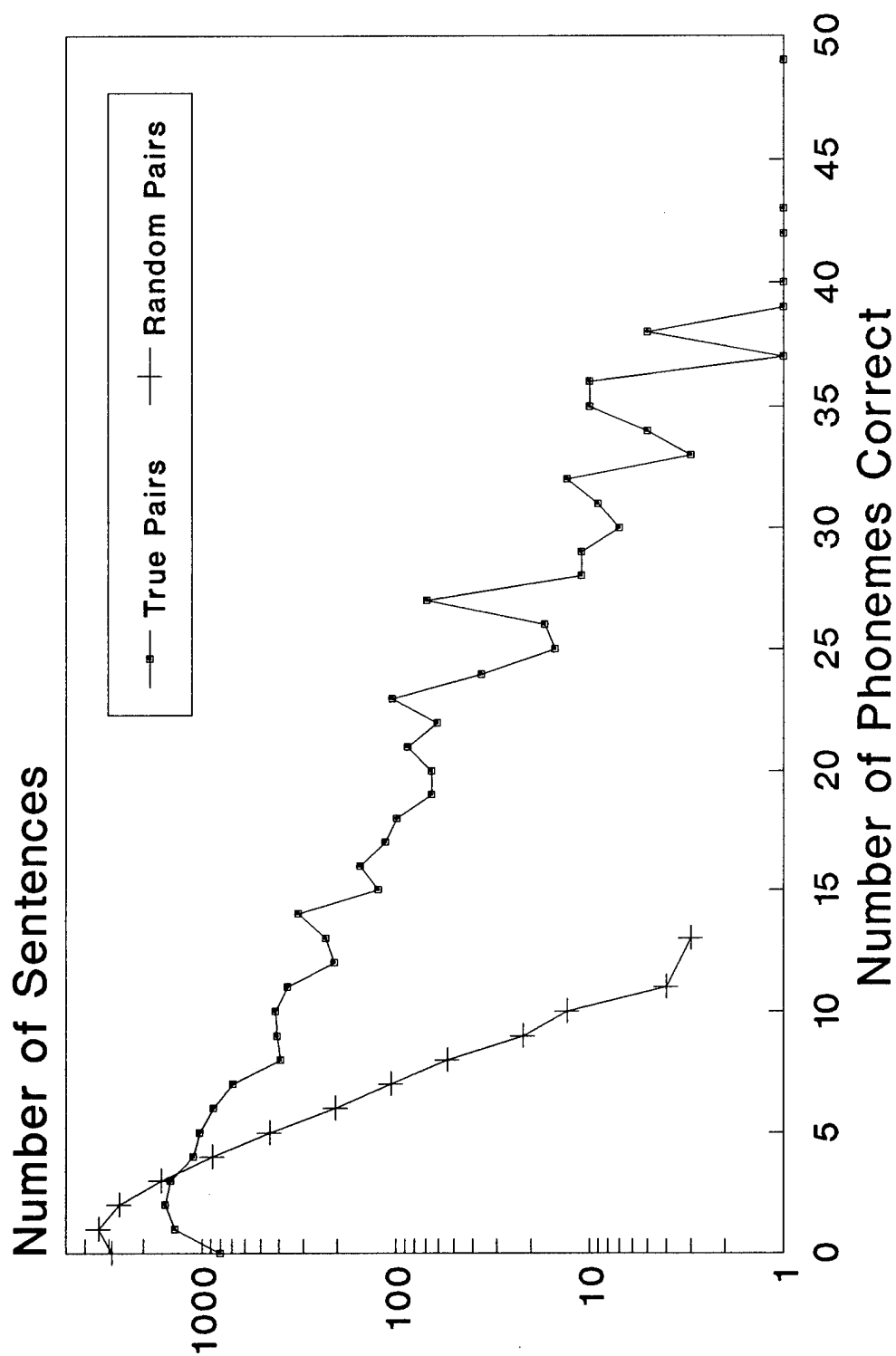


Figure 3. Distribution of number of phonemes correct as a function of the type of data: true versus random pairing of stimulus and response. N = 12,291 response sentences.

# Phoneme Substitution Uncertainty for True and Random Sentence Pairs

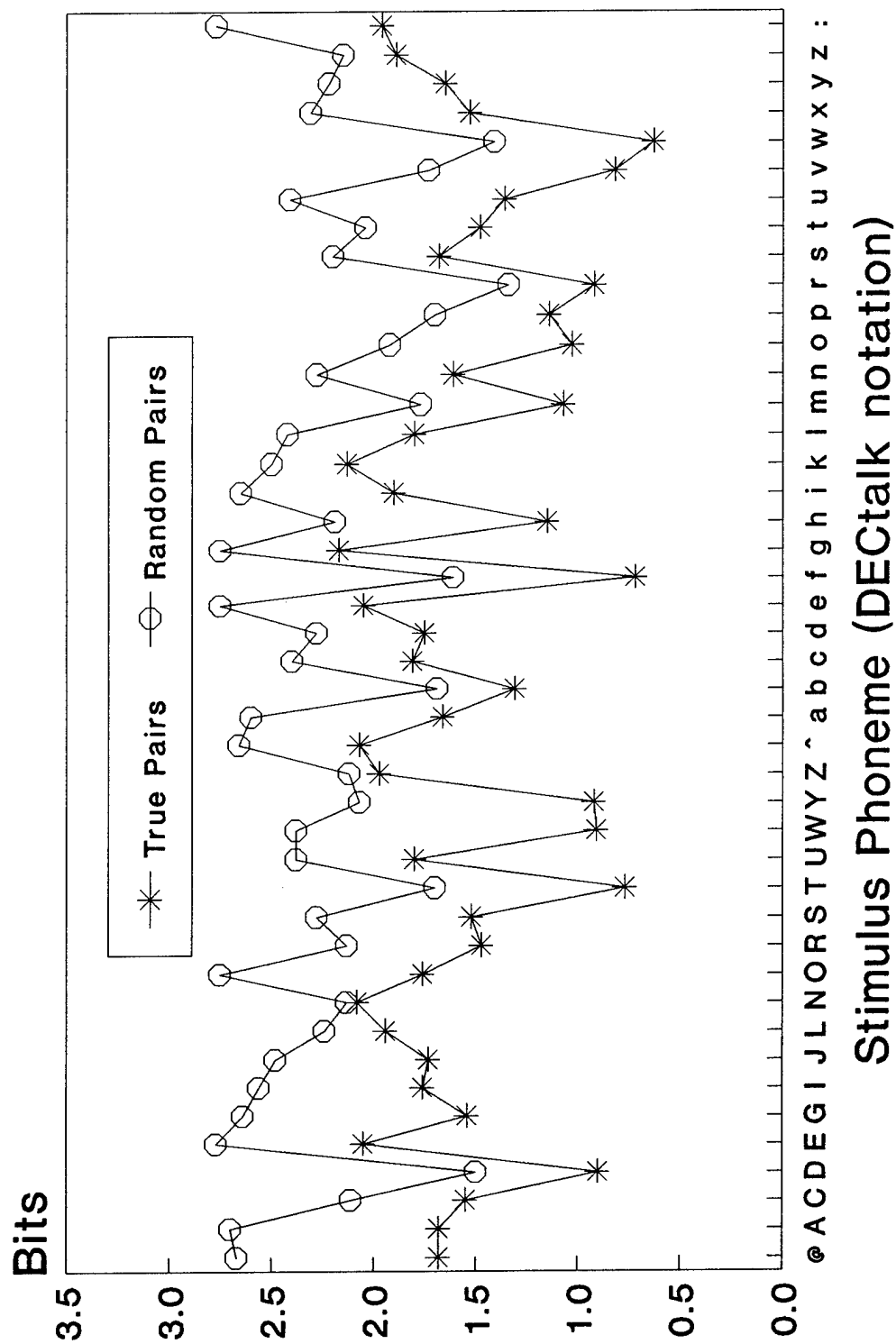


Figure 4. Average substitution uncertainty for each stimulus phoneme as a function of the type of data: true versus random pairings of stimulus and response. N = 12,291 response sentences.

## Feature Analysis on Consonant Substitutions in Alignments

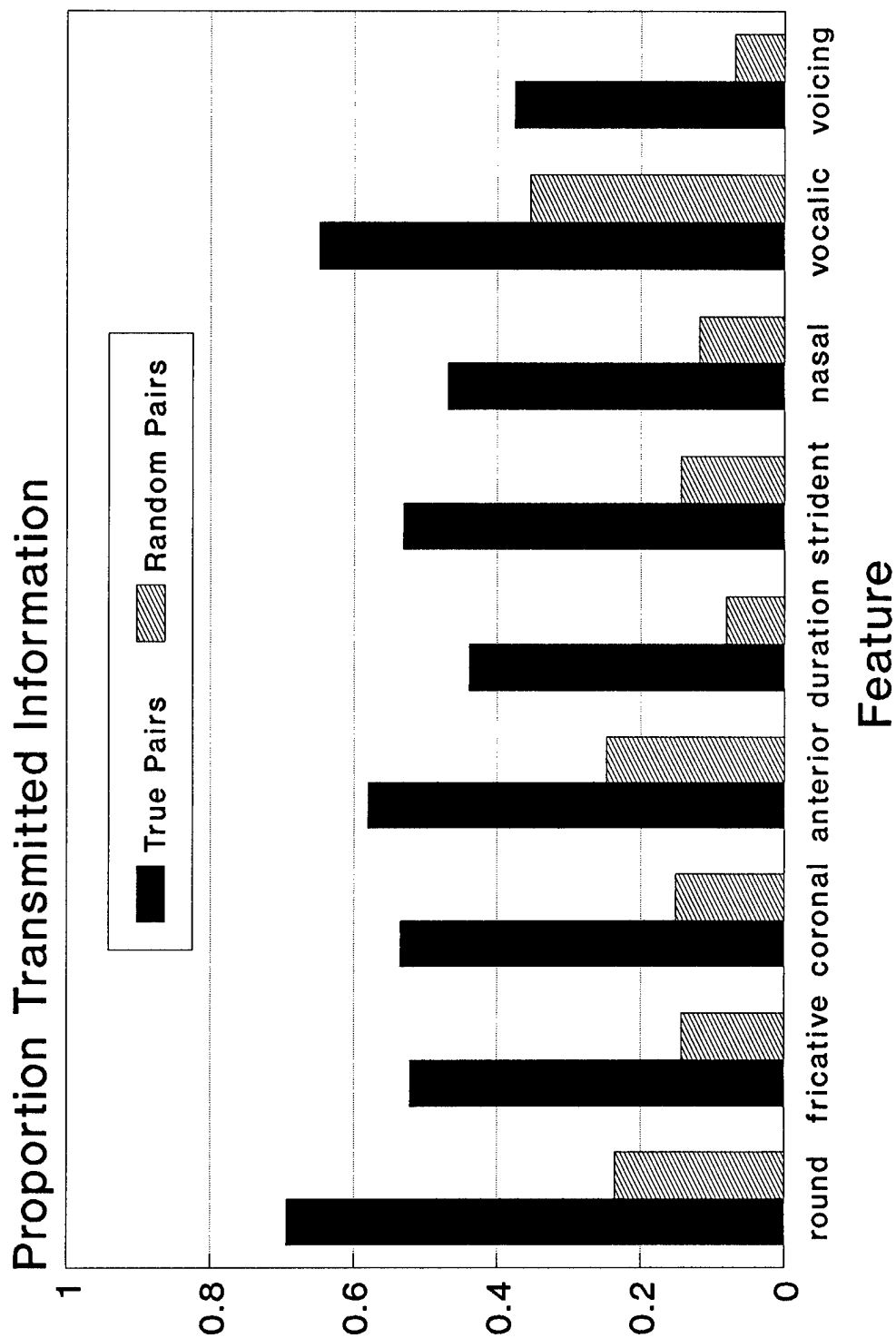


Figure 5. Feature analysis on consonant substitutions in sentence alignments as a function of the type of data: true versus random pairings of stimulus and response.



# Phonemes Correct In Sentences

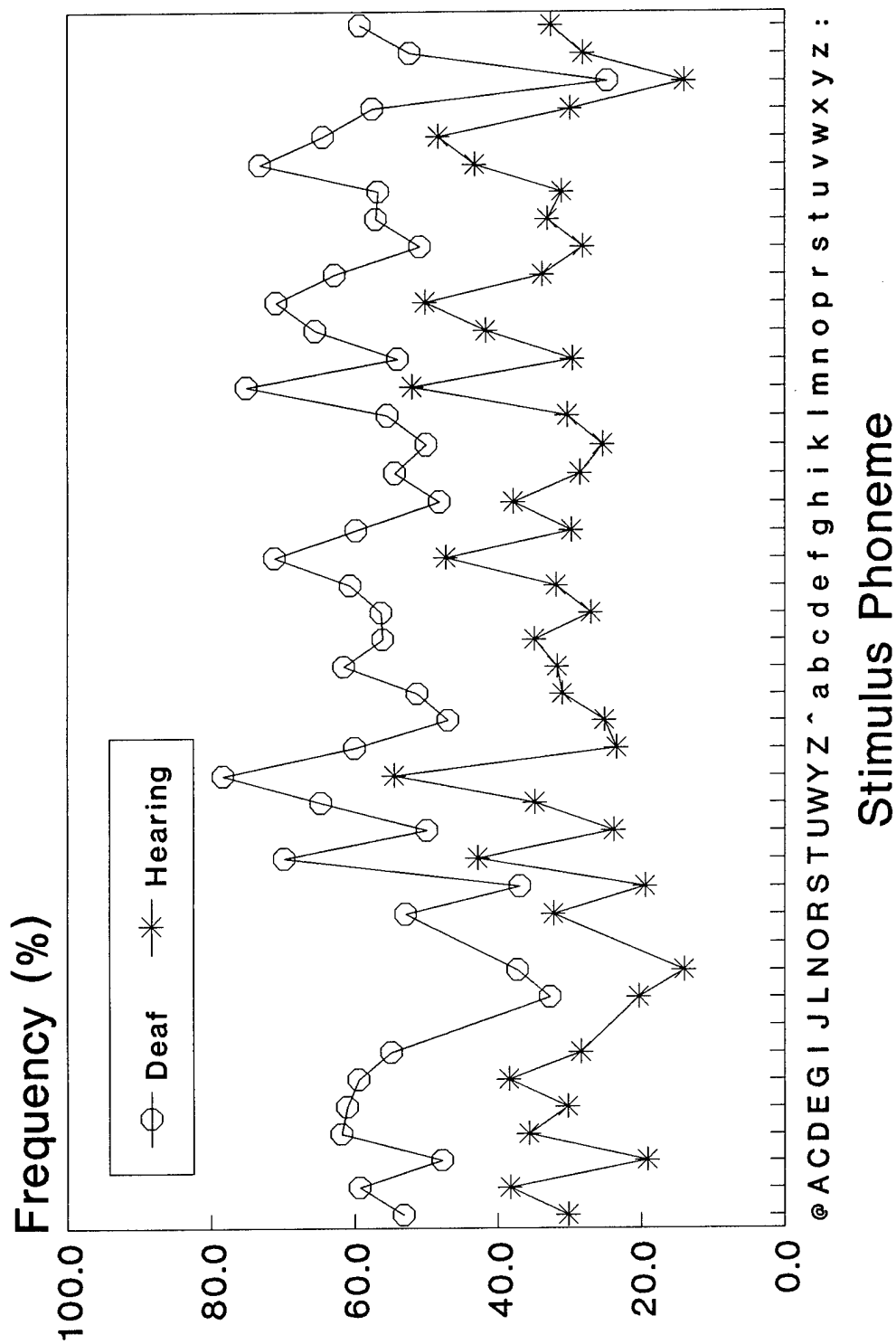


Figure 6. Frequency of phonemes correct as a function of subject group: deaf versus hearing. N = 7,554 response sentences.

# Phoneme Substitution Uncertainty In Sentences

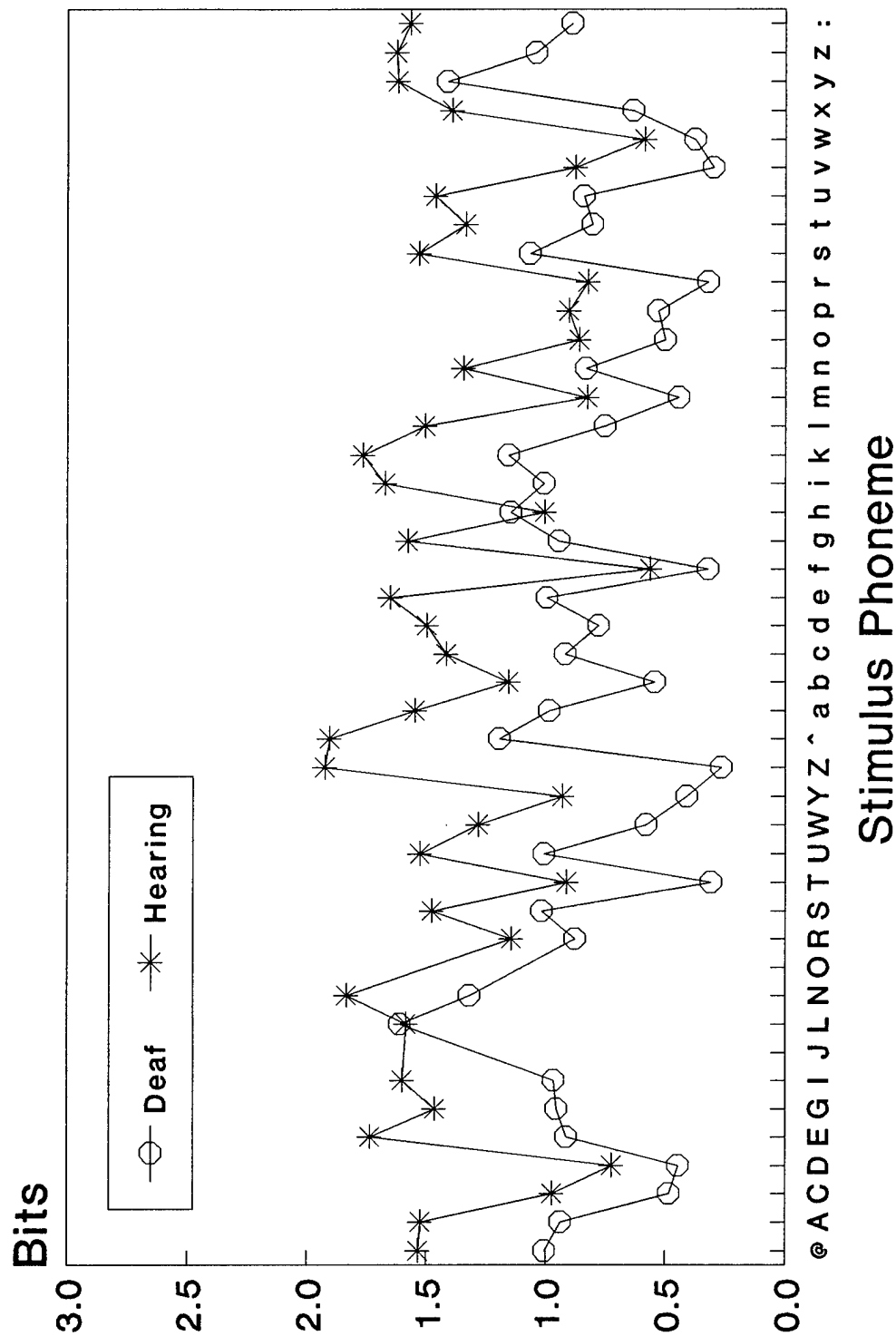


Figure 7. Average substitution uncertainty for each stimulus phoneme as a function of subject group: deaf versus hearing.

Phoneme confusion matrices were constructed by extracting the individual phoneme-to-phoneme alignments from the database of alignments for true and randomly paired sentences. An information measure, substitution uncertainty in bits, was calculated for each phoneme in each of the databases as:

$$- \sum p_k \log_2 p_k,$$

where  $p_k$  is the proportion of responses in category  $k$  and  $k$  is an index of summation that represents each possible substitution error. Figure 4 shows average phoneme substitution uncertainty for phonemes in true and random sentence pairs. Uncertainty is always higher for the phonemes in random pairs.

It was possible to examine also the confusion matrices in terms of more conventional transmitted information (TI) analyses using features. TI analysis uses the data in the entire confusion matrix, whereas substitution uncertainty considers only the off-diagonal entries. TI analysis was conducted using 12 features from Chomsky and Halle (1968) and two additional features, duration and frication, from Miller and Nicely (1955). Figure 5 shows proportion TI for the features that emerged as important when a sequential information analysis (Wang and Bilger, 1973) was applied to the two confusion matrices<sup>3</sup>. The figure shows that, in contrast with the data from alignments of true pairs, very little information was present in the confusion matrix from alignment of randomly assigned pairs. In summary, the results of the validation experiment suggest that the sequence comparator is sensitive to the nature of the data submitted to it.

## APPLICATION OF SEQUENCE COMPARISON TO A NORMATIVE STUDY OF LIPREADING IN DEAF VERSUS HEARING SUBJECTS

Next, an experiment was conducted to determine whether measures from the sequence comparator are sensitive to subject population differences (Bernstein et al., 1993b). A normative study was conducted in which 96 adult subjects with normal hearing and 72 adult subjects with profound hearing impairment lipread nonsense syllables, isolated words, and sentences. The sentence stimuli of interest here were 50 of the CID Everyday Sentences. The lipreading task was as described above. The data preparation was the same as described previously.

Figure 6 shows percent phonemes correct in sentences for the two groups. The figure shows that on average the deaf subjects were more accurate lipreaders than were the hearing subjects. This result is predictable, since percent words correct was approximately 20% for deaf and 40% for hearing subjects, a result that was obtained with the more conventional words correct scoring. The use of the sequence comparator does not actually contribute much to our understanding at this level of analysis, except to afford information about specific phonemes. The phoneme substitution uncertainty measure does contribute novel insight. Figure 7 shows phoneme substitution uncertainty obtained for each of the phonemes in alignments from the two subject groups. The figure shows that substitution uncertainty is higher for the hearing subjects for almost every phoneme. An interpretation of this result

<sup>3</sup> Note that the figure does not give conditional proportions TI, although selection of the features was based on sequential information analysis (Wang & Bilger, 1973).

## Feature Analysis on Consonant Substitutions in Alignments

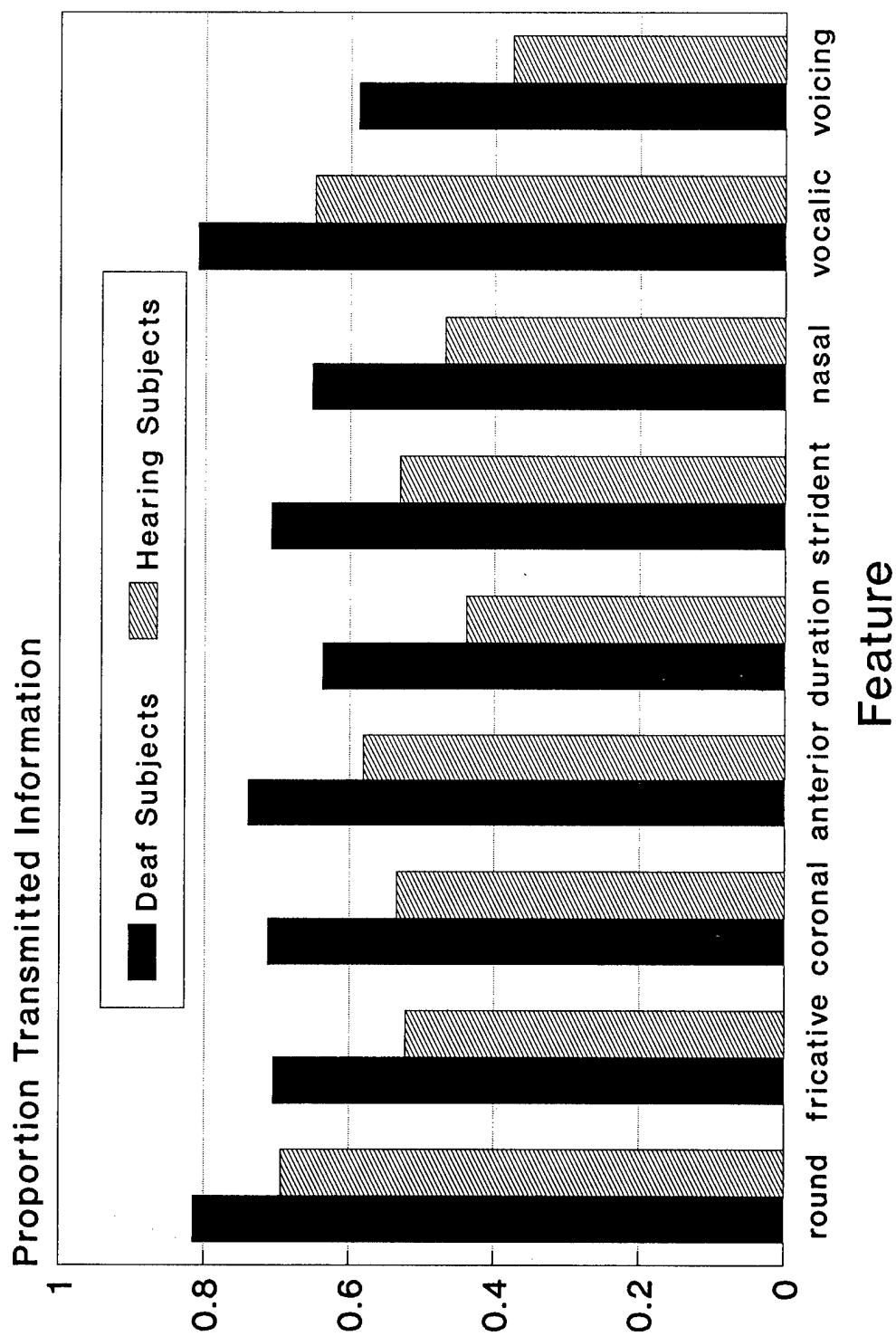


Figure 8. Feature analysis on consonant substitutions in sentence alignments as a function of subject group: deaf versus hearing.

# Feature Analysis on Nonsense Syllables and Substitutions in Alignments

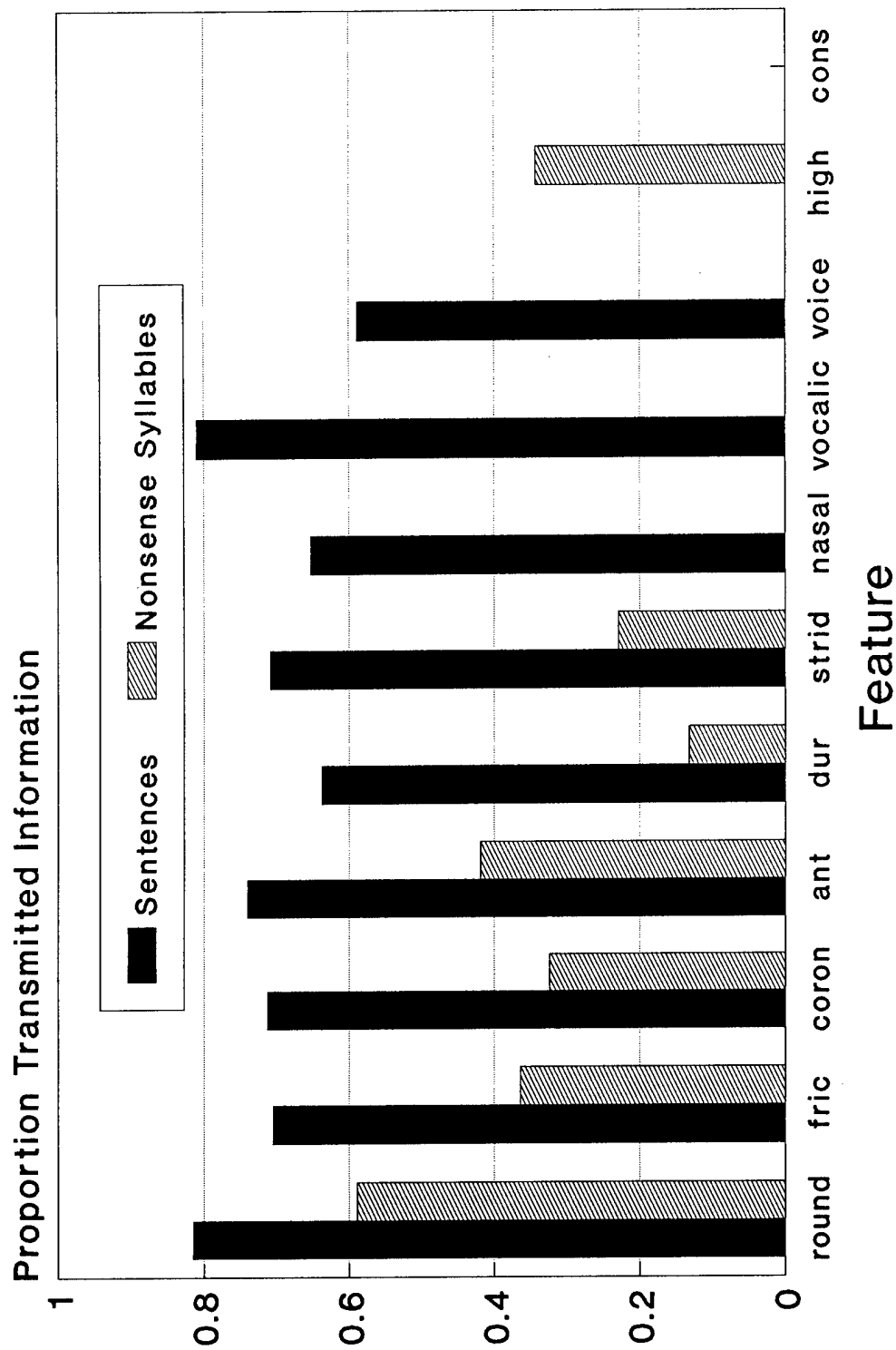


Figure 9a. Deaf Subjects

Figure 9a-b. Feature analyses on consonant substitutions in sentence alignments and on confusions in nonsense syllable identifications.

# Feature Analysis on Nonsense Syllables and Substitutions in Alignments

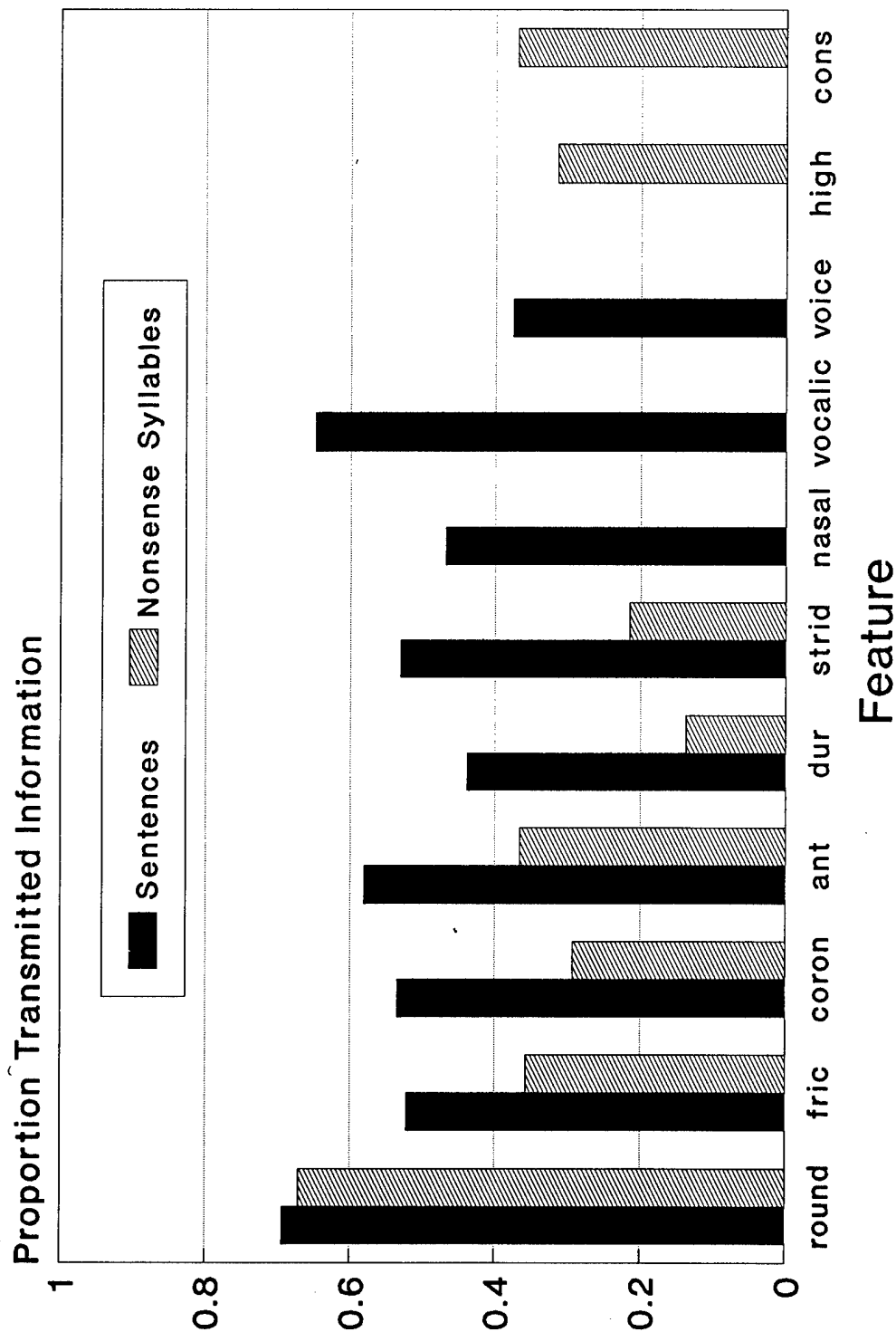


Figure 9b. Hearing Subjects

Figure 9a-b. Feature analyses on consonant substitutions in sentence alignments and on confusions in nonsense syllable identifications.

# CID Sentence Histogram #1

## "Walking's my favorite exercise"

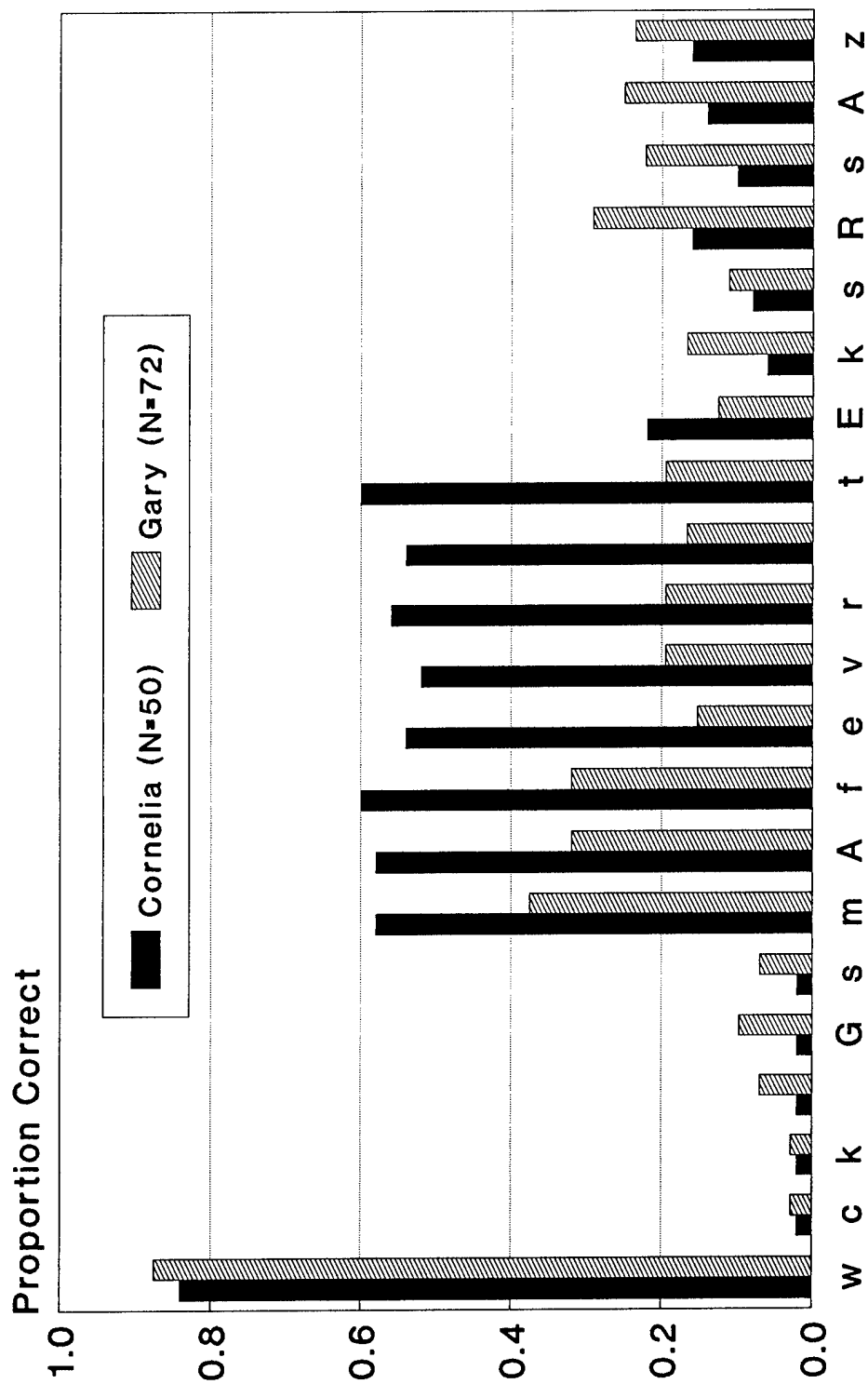


Figure 10. Sentence histogram: a response distribution for each stimulus phoneme in CID Everyday Sentence 1. Characters on the horizontal axis represent the DECTalk transcription of the sentence, Walking's my favorite exercise. Male talker = Gary, Female talker = Cornelia.

## Response

## Alignment

1	She won't let for a minute	#Si L#on- li#- bi#gcn#x#fy#- mIn ts# #Si#w ont#lE t#-- --- - fR#x#mIn t-#
2	You know you'll be gone for months	#Si L#o --n li#bi#gcn#x#fy#mIn ts# #-Y#n o#yuL#-- bi#gcn#- fR#m^n-Ts#
3	Don't be for me	#SiL#onl i#bi#gcn#x#fy#mIn ts# #--d ont#- bi#--- - fR#mi-----#
4	Don't look out for me	#SiL#on- li#bi#gcn#x#- fy#mIn ts# #--d ont#lU -- --k#W t#fR#mi-----#
5	You know who my offer please	#Si L#o nli#bi#gcn#x#fy#mIn ts# #-Y#n o#h-u#mA#--- c fR#p-li-z#
6	You don't even belong in this	#Si L#onl i#--- bi#gcn#x#fy#mIn - ts# #-Y#d ont#i vxn#bx lCG#- -- -In#DI-s#

Figure 11. Selected alignments for "She'll only be gone a few minutes."



is that substitutions (i.e., replacements only) made by deaf subjects are more systematic than those made by hearing subjects.

The more conventional TI analysis on the confusion matrices also showed a difference between subject groups. In Figure 8 it can be seen that the proportion TI was higher for the deaf subjects for each of the features.

Since subjects had also performed forced-choice CV nonsense syllable identification for 22 initial consonants, it was possible to compare TI across stimulus materials. Figure 9a-b shows results for nonsense syllables versus sentences for each of the subject groups. Note that the sentence data are the same as in Figure 8. Figure 9a-b shows that deaf subjects were more successful identifying nonsense syllables. Of more interest, however, is that 1) higher levels of TI were obtained with sentence materials than with nonsense syllables, and 2) somewhat different features emerged as important with the two different types of stimulus materials. The features high and consonantal were only important for nonsense syllable identification. The features nasal, vocalic, and voicing were only important for sentence identification. Since the visual phonetic stimulus does not afford all the featural distinctions (such as voicing), presence of these features in the sentence data can be attributed to the recognition of words.

In summary, the sequence comparator was shown to be sensitive to subject group differences. Substitution uncertainty measures suggest there is a qualitative difference between subject populations. The capability to extract confusion matrices from alignments was shown useful in comparing TI analyses for nonsense syllable identification versus open set sentence identification.

## **OTHER USES FOR ALIGNMENTS**

### **Sentence Histograms**

Another use for alignments appears in a paper by Demorest and Bernstein (1991). They introduce the sentence histogram, a figure showing performance accuracy on a phoneme-by-phoneme basis throughout a sentence (see Figure 9). Although we have not examined such histograms formally in great detail, informal study suggests that the perceptual process of visual speech perception may involve attempting to spot words that may become salient in the context of otherwise ambiguous or unintelligible speech. The reason for this hypothesis is that we have observed islands of correct word identifications embedded in otherwise incorrect responses (as in Figure 10). This hypothesis contradicts a common explanation for how lipreading is accomplished. That is, the lipreader is said to use context to recover unintelligible words. This explanation cannot account for identification of words in otherwise unintelligible contexts. Our observations have led to the hypothesis that lipreading is data driven to a far greater extent than has heretofore been thought.

### **Word boundary detection**

A different phenomenon that can be observed in our data is failure to detect word boundaries. Consider the alignments in Figure 11 for responses to the stimulus, She'll only be gone a few minutes.

These alignments were chosen because they provide evidence for systematic word boundary errors that may reflect effects of normal processes that contribute to segmentation.

In the six examples, the /L/ in the stimulus word she'll is aligned with the word-initial consonant of a response word. It appears that the more prominently stressed /o/ in only has "captured" the preceding consonant. Later in the sentence, correct parsing is reestablished, likely due to the high visibility of be.

Recently, Cutler and Butterfield (1992) described an experiment in which subjects listened to connected speech with controlled stress rhythm of strong and weak syllables, controlled lexical stress in terms of the location of stressed syllables within multisyllabic words, and controlled phonetic length of vowels. Subjects reported what they heard. Because the stimuli carefully controlled prosodic factors, it was possible to investigate systematic patterns of boundary shift errors hypothesized to be due to prosody. However, relatively few of the obtained responses could be evaluated, because the investigators did not have a method to make use of partial responses. Responses that did not explicitly reflect the number of syllables and the rhythmic pattern in the stimulus were rejected, and only 42% of the responses satisfied criteria. Nevertheless, support was obtained for the hypothesis that listeners use prosodic information to parse word boundaries. With sequence comparison, this type of interesting hypothesis could be more efficiently and elaborately investigated.

## **Conclusions**

The experiments reported here demonstrate that sequence comparison can be applied to research on perception of connected speech. The sequence comparator produces several different numerical measures such as number of phonemes correct, number of insertions, deletions, and substitutions, and also phoneme-to-phoneme alignments. Alignments can be submitted to further analysis in which patterns of response are extracted. Although the data discussed here were from lipreading experiments, sequence comparison techniques can be applied to auditory and audio-visual connected speech as well as to visual speech perception.

## **ACKNOWLEDGEMENTS**

This research was supported by the following NIH grants: DC00023, NS22183, and DC00695.

## **REFERENCES**

- Bernstein, L.E., Demorest, M.E., and Eberhardt, S.P. "A computational approach to analyzing sentential speech perception: Phoneme-to-phoneme stimulus-response alignment" (1993a, submitted).
- Bernstein, L.E., Demorest, M.E., and Tucker, P.E. Speech Perception Does Not Require Audition: Evidence from Profoundly Hearing-Impaired Speechreaders (1993b, in preparation).

Chomsky, N., and Halle, M. The Sound Pattern of English, New York: Harper and Row (1968).

Cutler, A., and Butterfield, S. "Rhythmic cues to speech segmentation: Evidence from juncture misperception," *J. Mem. Lang.*, 31, 218-236 (1992).

Davis, H., and Silverman, S.R. Hearing and Deafness, New York: Holt, Rinehart, and Winston (1970).

Demorest, M.E., and Bernstein, L.E. "Computational explorations of speechreading," *J. Acad. Reh. Aud.*, 24, 97-111 (1991).

DECtalk DTC01 Programmer Reference Manual, Educational Services Department, Digital Equipment Corporation, Maynard MA (1984).

Fisher, C.G. "Confusions among visually perceived consonants," *J. Speech Hear. Res.*, 11, 796-804 (1968).

Kruskal, J.B. "An overview of sequence comparison," in Time warps, string edits, and macromolecules: The theory and practice of sequence comparison, edited by D. Sankoff and J.B. Kruskal, Reading MA: Addison-Wesley (1983).

Miller, G.A., and Nicely, P.E. "An analysis of perceptual confusions among some English consonants," *J. Acoust. Soc. Am.*, 27, 338-352 (1955).

Owens, E., and Blazek, B. "Visemes observed by hearing-impaired and normal-hearing adult viewers," *J. Speech Hear. Res.*, 28, 381-393 (1985).

Sankoff, D., and Kruskal, J.B. Time warps, string edits, and macromolecules: The theory and practice of sequence comparison, Reading MA: Addison-Wesley (1983).

Wang, M.D., and Bilger, R.C. "Consonant confusions in noise: A study of perceptual features," *J. Acoust. Soc. Am.*, 54, 1248-1266 (1973).

# APPLICATIONS OF GENERALIZABILITY THEORY TO MEASUREMENT OF INDIVIDUAL DIFFERENCES IN SPEECH PERCEPTION<sup>1</sup>

Marilyn E. Demorest

University of Maryland  
Baltimore County Campus  
Baltimore MD 21228

Assessment of individual differences in speech perception requires standardized tests that are sensitive to the relevant sources of variability in test scores (i.e., valid) and insensitive to irrelevant, extraneous sources of variability (i.e., reliable). Test reliability has traditionally been evaluated by examining extraneous sources of variability independently of one another. For example, retest reliability evaluates the consistency of test scores over time, with test occasion being the extraneous variable. Alternate-form reliability evaluates the consistency of scores over different test forms, with test form being the extraneous variable. Split-half reliability and internal consistency reliability evaluate consistency of performance over items within a single test form, and interscorer reliability reflects consistency across scorers.

Generalizability theory (Cronbach, Gleser, Nanda, and Rajaratnam, 1972) is a statistical theory of sources of variability in behavioral observations that permits estimation of the effects of several extraneous variables, and their interactions, within a single experiment. A generalizability study is an experiment in which potential sources of variability in test scores are manipulated. A statistical model for a single observation and an analysis-of-variance model appropriate for the experimental design are specified. Next, expected values of the mean squares from the analysis of variance are determined and used to estimate the variance component for each source of variability in the observations.

As an example, consider a study conducted by Demorest and Cord (1993) in which four monosyllabic word lists (NU-6) were administered on each of two days to a sample of 40 hearing-impaired adults. The sources of variability were the test list and the test occasion. The statistical model for the score of one subject on a given list on a given day is:

$$X = \mu + \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 + \alpha_5 + \alpha_6 + \epsilon,$$

where  $\mu$  is a grand mean, the  $\alpha$  parameters represent the effects of Subject, List, Day, List x Day, Subject x List, and Subject x Day, and  $\epsilon$  is random, residual error. Given this model for a single score, the variance of observed scores is:

---

<sup>1</sup>Based on a paper submitted to the Journal of the Academy of Rehabilitative Audiology. This work was supported in part by NIH grant DC00695.

$$\sigma_x^2 = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2 + \sigma_5^2 + \sigma_6^2 + \sigma_e^2$$

The goal of the generalizability analysis is to estimate each of these variance components and their contribution to the total observed variance. Of all of these sources, only the first, that for Subject, reflects relevant variance; all other components are extraneous to the purposes of testing.

The expected value of each mean square in the analysis of variance is a linear combination of the variance components. By equating each mean square to its expected value, estimates of the variance components can be obtained. For example, the expected value of the mean square for the interaction of Subject x Day,  $MS_6$ , is  $4\sigma_6^2 + \sigma_e^2$ , whereas the expected value of the mean square for residual error,  $MS_E$ , is  $\sigma_e^2$ . Thus  $[MS_6 - MS_E]/4$  provides an estimate of  $\sigma_6^2$ . An algorithm for deriving the expected values of mean squares is given in Winer (1971).

Analysis of the data from Demorest and Cord (1993) produces the following estimates for each component and for the total variance:

$$\hat{\sigma}_x^2 = \hat{\sigma}_1^2 + \hat{\sigma}_2^2 + \hat{\sigma}_3^2 + \hat{\sigma}_4^2 + \hat{\sigma}_5^2 + \hat{\sigma}_6^2 + \hat{\sigma}_e^2$$

$$70.74 = 57.34 + .32 + 0 + 0 + 1.52 + 4.84 + 6.73$$

The largest source of variability is Subject, accounting for 81.1% of the total variance of observed scores. Although List is a statistically significant (i.e., non-zero) effect, its magnitude is quite small. Effects for Day and the interaction of List x Day produce negative variance estimates, which have been set to zero. The interactions of Subject x List and Subject x Day, and the residual error (which represents the combined effects of all other sources of variability) account for 2.1%, 6.8%, and 9.5% of the variance, respectively.

Bilger, Nuetzel, Rabinowitz, and Rzechkowski (1984) performed a generalizability analysis of the Speech Perception in Noise (SPIN) test in which several variables were manipulated. Given the large number of subjects in their study ( $N = 128$ ), the large number of lists (10), and the multiple conditions of testing/scoring, they found that many irrelevant sources of variability were statistically significant. Their variance component analysis, however, revealed that the magnitude of many of these effects was trivial. Of particular importance was the finding that differences between methods of scoring (immediate write-down by the examiner versus transcription from a recording of the subject's response) were virtually zero.

Generalizability analysis has also been used by Demorest, Bernstein, and Tucker (1993) to compare speechreading performance in two populations of subjects. Normal-hearing subjects ( $N = 96$ ) and hearing-impaired subjects ( $N=72$ ) speechread 50 video-recorded CID Everyday Sentences, half spoken by a female talker and half spoken by a male talker. The unit of observation was the number of words correct on a single sentence. The effect of Group (normal vs. impaired hearing) accounted for 7.4% of the variance in scores, while individual differences among subjects within the groups accounted for 19.0%. Individual test items accounted for 18.5% and residual error 51.1%. The remaining variance components combined accounted for only 4% of the total variance. The latter finding is important because it suggests that the interaction of Group x Item is a small effect. Thus, in

developing speechreading materials (tests) for hearing-impaired individuals, it would be possible to use subjects with normal hearing because the relative difficulty of items is very similar for the two groups. Likewise, absence of a Group x Talker interaction implies that talker differences are the same within each population.

When data for the hearing-impaired and normal-hearing groups were analyzed separately, the variance component for Subject was more than twice as large in the hearing-impaired group (2.14 vs. 0.99). Differences among test items were more than three times as great (2.75 vs. 0.81 ). It appears that the wider range of speechreading ability in the hearing-impaired sample was also reflected in the mean performance on individual items. Other variance components, however, were very similar for the two groups. Variance attributable to Talker was essentially zero, as was the interaction of Subject x Talker, and residual error variances were 4.67 and 3.49 for the hearing-impaired and normal-hearing groups, respectively.

Generalizability analysis yields coefficients of generalizability, which are analogous to reliability coefficients. Each coefficient is based on a data collection model for testing and a universe of generalization for test score interpretation. Together these determine which sources of variability affect observed scores and universe scores. (The latter are analogous to true scores in classical test theory.) The coefficient equals the ratio of universe-score variance to observed-score variance. For example, given the data from Demorest and Cord (1993) on NU-6 word lists, we might specify a data collection model as administration of a single list to a subject on a given day. Table 1 shows the estimated generalizability coefficient for four universes of generalization. Also shown are average reliability coefficients obtained from the same data. For example, the generalizability coefficient for generalization across lists, but not days, is analogous to an alternate-form reliability coefficient. The average of all the alternate-form reliability coefficients in these data gives a value virtually identical to the generalizability coefficient. Generalizability theory also makes it possible, however, to estimate immediate retest reliability (same list, same day,  $r = .904$ ), even though no immediate retests were given.

Table 1. Estimated Generalizability Coefficients for Four Universes of Generalization and Analogous Reliability Coefficients from Demorest and Cord (1993).

Universe of Generalization	Estimated Generalizability Coefficient	Mean Observed Reliability Coefficient
Across Lists and Days	.814	.808
Across Lists for a Given Day	.883	.877
Across Days for a Given List	.836	.832
None: A Given List on a Given Day	.904	

Generalizability theory is especially useful for estimating the number of test items needed to achieve a particular level of generalizability. For example, Demorest and Bernstein (1992) presented speechreading data on 104 subjects with normal hearing who viewed 100 video-recorded CID Everyday Sentences, 50 for each of two talkers. The unit of observation was the subject's score on a single sentence scored in terms of total words correct. Generalizability coefficients were estimated for three models of data collection and generalization:

Model 1: Test with a single talker, generalize over all test items by this talker.

Model 2: Test with a single talker; generalize over all test items and both talkers.

Model 3: Test some subjects with one talker, others with the other talker; generalize over all test items and both talkers.

As can be seen in Figure 1 (adapted from Demorest and Bernstein, 1992), generalizability is highest for Model 1 and lowest for Model 3. All three functions, however, begin to plateau at about 30-40 items, suggesting that for these recordings of these materials, individual differences among subjects can be reliably estimated with about 40 sentences.

Generalizability theory provides an integrated framework for evaluating multiple sources of variability in behavioral observations and for deriving implications for test development and test score interpretation. It has only recently begun to be applied in the domain of speech perception, but as the examples presented here illustrate, it can provide valuable insights about individual differences, both within and between normal-hearing and hearing-impaired populations.

## REFERENCES

- Bilger, R.C., Nuetzel, J., Rabinowitz, W.M., and Rzeczkowski, C. "Standardization of a test of speech perception in noise," *J. Speech and Hear. Res.*, 27, 32-48 (1984).
- Cronbach, L.J., Gleser, G.C., Nanda, H., and Rajaratnam, N. The dependability of behavioral measurements: Theory of generalizability for scores and profile. New York: Wiley (1972).
- Demorest, M.E., and Bernstein, L.E. "Sources of variability in speechreading sentences: A generalizability analysis," *J. Speech and Hear. Res.*, 35, 876-891 (1992).
- Demorest, M.E., and Bernstein, L.E. "Applications of generalizability theory to measurement of individual differences in speech perception," *J. Acad. of Rehab. Audiology*, 26 (1993).
- Demorest, M.E., Bernstein, L.E., and Tucker, P.E. "Sources of variability in speechreading nonsense syllables, isolated words, and sentences for subjects with hearing impairment," manuscript in preparation (1993).

# Speechreading CID Sentences

## Demorest & Bernstein (1992)

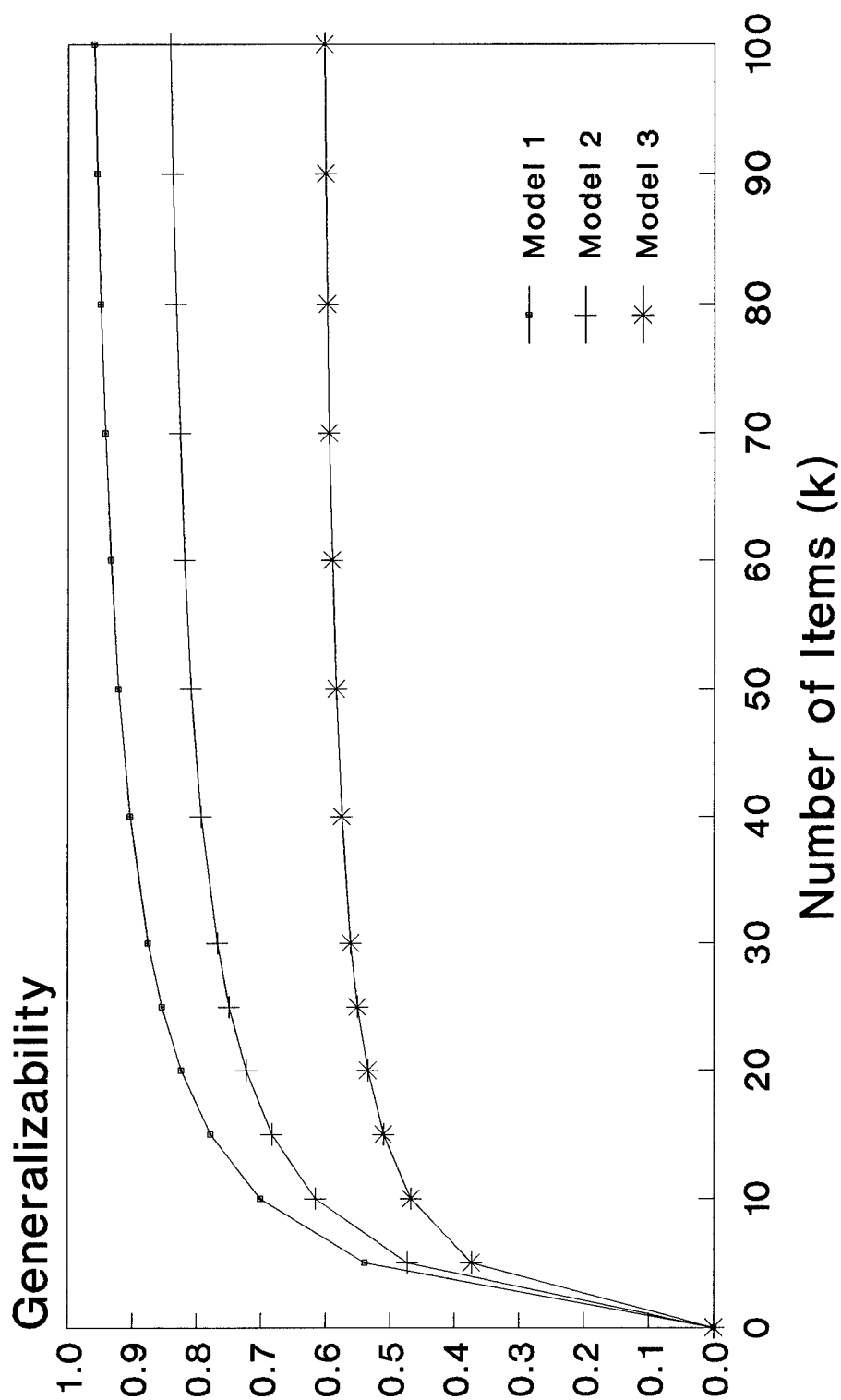


Figure 1  
127



Demorest, M.E., and Cord, M. "Evaluation of temporal and interlist sources of variability in NU-6 test scores," manuscript in preparation (1993).

Winer, B.J. Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill (1971).

# INTEGRATION MODELS OF SPEECH INTELLIGIBILITY

Louis D. Braida

Research Laboratory of Electronics  
Massachusetts Institute of Technology  
Cambridge MA 02139

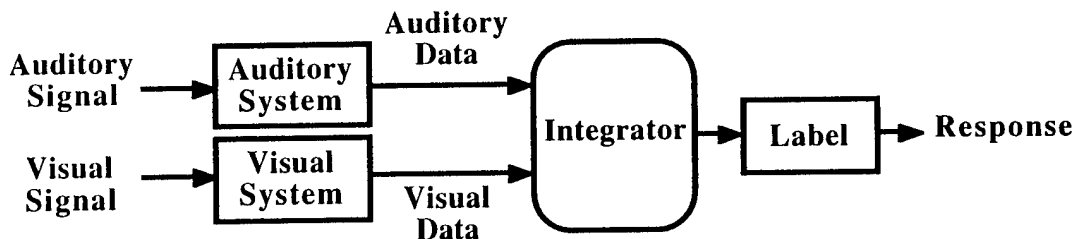
***Abstract.*** This paper reviews current research concerned with modelling the effects of cue integration on speech intelligibility, as in the case of audiovisual speech reception. Two "optimum-processor" models have been applied to predict performance when cues are integrated on the basis of performance measures obtained in conditions when only partial cues are available. In Pre-Labeling Integration, continuous sensory data are combined from different cue sources before response labels are assigned. In Post-Labeling Integration, the responses that would be made on the basis of each cue source are combined and a joint response is derived from the pair. To describe Pre-Labeling Integration, confusion matrices are characterized by a multidimensional decision model that allows performance to be described by a subject's sensitivity and bias in using continuous-valued cues. The cue space is characterized by the locations of stimulus and response centers. The distance between a pair of stimulus centers determines how well two stimuli can be distinguished in a given experiment. These models have been shown to provide relatively accurate accounts of the integration of auditory and visual cues in identification experiments in which the stimuli are syllables distinguished by consonant content. In this paper the integration models are used to relate performance in high-pass/low-pass filtering conditions to the wideband condition and to predict the changes in performance associated with changes in signal-to-noise ratio.

## INTRODUCTION

In the reception of speech, cues derived from several sources are integrated in ascertaining the message intended by the talker. For example, by integrating cues derived from low-frequency spectral regions that are highly informative about the voicing and nasality of consonants, with cues derived from high-frequency spectral regions that are highly informative about place of production, it is possible to specify the consonant spoken precisely. Similarly, when the acoustic speech waveform is degraded by noise or low-pass filtering, but the face of the talker is visible, the spoken message can be determined by integrating acoustic and visual cues. At the present time our understanding of the processes by which cues are combined in the process of speech reception is still fairly tentative. In this section we discuss models of the integration process that can be used to understand how well listeners are able to integrate such cues. Our initial efforts were focused on the problem of multimodal integration for consonant segments (Braida, 1991). Currently we are attempting to extend this work to the problem of integrating across spectral regions in acoustic speech reception.

In English, reception of consonants is an important determinant of intelligibility and there are extensive data on auditory, visual, and audiovisual consonant segment reception. In the following paragraphs we describe two "optimal processing" models for the types of integration that can be used when more than one source of cues is available (Fig. 1). In Pre-Labeling Integration, continuous sensory data are combined across modalities before response labels are assigned. In Post-Labeling Integration, the responses that would be made under unimodal conditions are combined and a joint response is derived from the pair.

## Pre-Labeling Integration



## Post-Labeling Integration

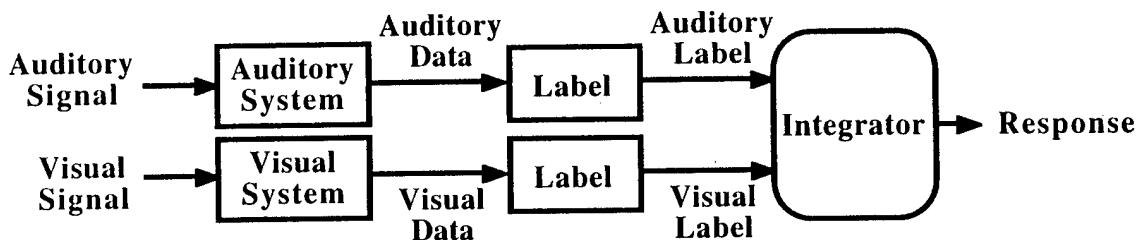


Figure 1: Two "optimum-processor" models of audiovisual integration. In Pre-Labeling Integration sensory data are combined across modalities before a response label is assigned. In Post-Labeling Integration the responses that would have been assigned to the visual stimulus and the auditory stimulus are used to derive a joint response to the audiovisual stimulus.

### Pre-Labeling Integration

In Pre-Labeling Integration, consonant reception is described in terms of a multidimensional extension of the theory of signal detection (Braid, 1988). Consonants are assumed to be identified on the basis of a noisy vector of cues  $\vec{X}$  (Fig. 2). From presentation to presentation of consonant  $S_j$ , the cue vector  $\vec{X}$  is displaced from the *stimulus center*  $\vec{S}_j$  for that consonant by an additive noise vector whose components are independent Gaussian random variables with zero means and common unit variance. Corresponding to each response there is also a *response center* or *prototype*. The subject is assumed to respond  $R_k$  if and only if the distance from the observed vector of cues  $\vec{X}$  to  $\vec{R}_k$  is smaller

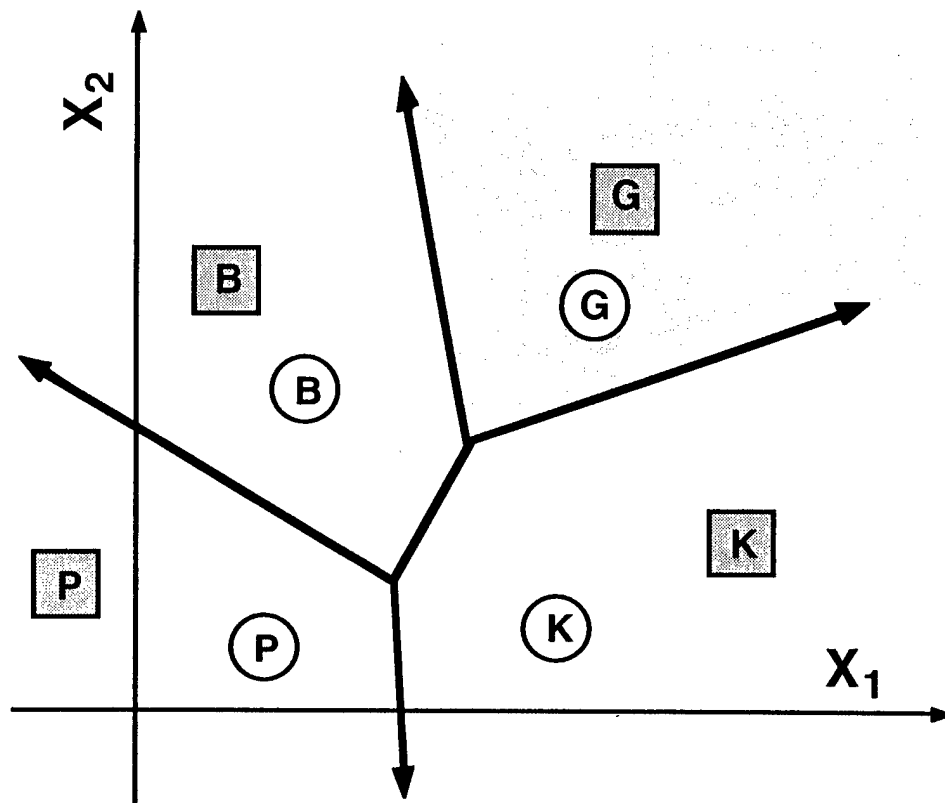


Figure 2: Two-dimensional cue space for an identification experiment with the four consonants /p, b, k, g/. Location of the stimulus centers (circles) relative to response prototypes (squares) is arbitrary in this figure. Response regions are bounded by line segments (extended bold lines), each of which coincides with a segment of the perpendicular bisector of a line connecting two response prototypes. All the points in a given response region are closer to the prototype in that region than to any other prototype. In general,  $P(R_j|S_i)$  is computed by integrating the multidimensional Gaussian density with mean  $\bar{S}_i$  over the response region that contains  $\bar{R}_j$ .

than the distance to any other prototype. In effect, this rule partitions the  $D$ -dimensional space of cue vectors into  $N$  compact "response regions" that are bounded by hyperplanes. All points in each region are closer to the prototype in that region than to any other prototype.

In general, the distance between two stimulus centers  $d'(i,j)$  determines an observer's ability to distinguish between the two consonants  $S_i$  and  $S_j$ , although the accuracy of identifying the stimuli can be reduced if there is response bias, i.e., if the response prototypes are poorly located relative to the stimulus centers.

When there are multiple sources of cues, we model the integration process by assuming that the cue densities in the multimodal condition are the Cartesian products of the densities corresponding to the separate modalities. This space is constructed from the orthogonal composition of the cue

spaces corresponding to each of the cue sources. Thus we assume that cues are combined optimally and that there is no perceptual interference (e.g., masking or distraction) across cue sources.

Consider, for example, a multimodal AV condition (Fig. 3). Locations of the stimulus centers in the combined cue space for three hypothetical identification experiments involving the consonants /p, b, k, g/ are shown. For purposes of illustration, we assume that confusions in the A and V conditions can satisfactorily be accounted for in 1-dimensional cue spaces. The cue space in the AV condition is therefore predicted to be 2-dimensional. This model for the combined condition predicts that there is a simple Pythagorean relation between a subject's sensitivity in distinguishing  $S_i$  from  $S_j$  in the multimodal condition,  $d'_{AV}(i, j)$ , and the corresponding unimodal sensitivities  $d'_A(i, j)$  and  $d'_V(i, j)$ :

$$d'_{AV}(i, j) = \sqrt{d'_A(i, j)^2 + d'_V(i, j)^2} \quad (1)$$

The geometric properties of the multimodal cue space capture the common observation that consonants can be distinguished in the AV condition if they can be distinguished in either (or both) of the A or V conditions. We have made predictions for multimodal accuracy in the special case in which the response centers coincide with the stimulus centers, i.e.,  $\vec{R}_i = \vec{S}_i$ .

### Post-Labeling Integration

In Post-Labeling Integration, the listener is assumed to process the cues from each source separately, to form tentative responses appropriate for each source, and to derive a joint response from the combination of individual responses. For example, in the AV case, presentation of stimulus  $S_i$  generates a pair of labels  $(A_m, V_n)$  corresponding respectively to the auditory and visual judgments. The labels correspond to the responses that would be given when only one cue source is available. The collection of label pairs is assumed to be divided into mutually exclusive and collectively exhaustive sets, where the composition of each set is chosen to maximize the probability of identifying the consonant correctly.

Since the label produced for each cue source is independent of the labels in the other sources, labelling judgments for each source are made by the same processes that operate in single source conditions. Thus the probability that the label pair  $(A_m, V_n)$  is elicited by stimulus  $S_i$  is given by

$$P^*_{AV}(A_m, V_n | S_i) = P^*_A(A_m | S_i) \times P^*_V(V_n | S_i) = P_A(R_m | S_i) \times P_V(R_n | S_i) \quad (2)$$

where  $P_A(R_m | S_i)$  and  $P_V(R_n | S_i)$  determine the frequencies observed in the auditory and visual unimodal confusion matrices.

When the *a priori* stimulus presentation probabilities of the stimuli are equal, the highest identification accuracy (probability of responding correctly) results from using the *maximum likelihood* rule: the response associated with the label pair  $(A_m, V_n)$  should be the identity of the stimulus for which  $P^*_{AV}(A_m, V_n | S_i)$  is greatest. If  $\Theta_j$  denotes the set of  $(m, n)$  for which  $P^*_{AV}(A_m, V_n | S_j) > P^*_{AV}(A_m, V_n | S_i)$  for all  $i \neq j$ , then the conditional response probabilities are predicted to be

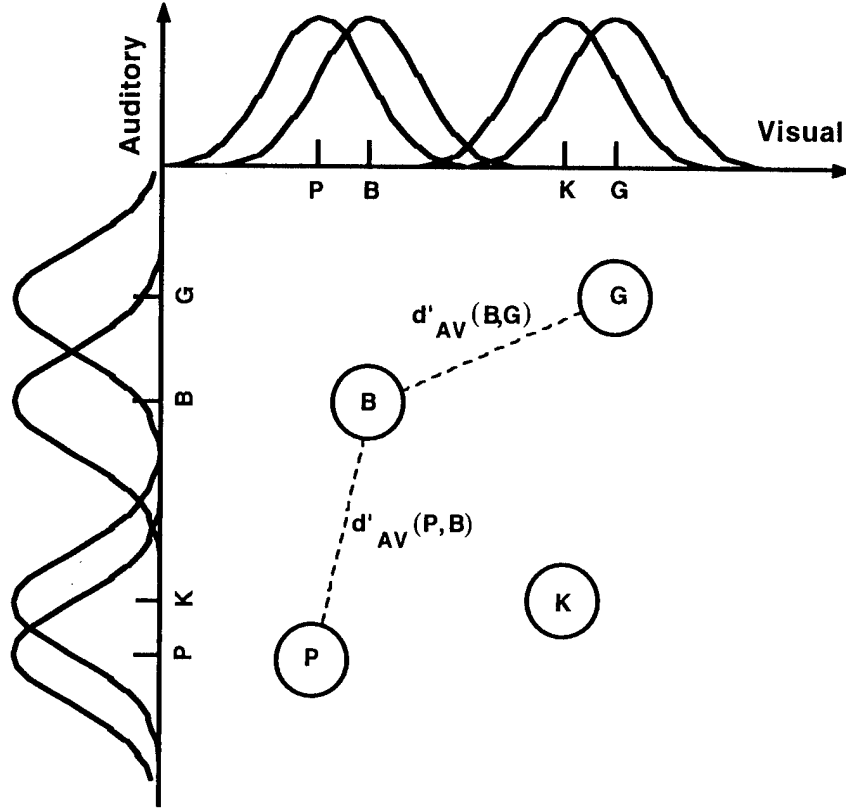


Figure 3: Hypothetical auditory, visual, and audiovisual cue spaces for /p, b, k, g/ identification experiments. Confusions in the auditory and visual conditions are each assumed to be adequately described by scalar cues. Stimulus densities for these scalar cues are shown on the  $A$  and  $V$  axes scaled to have a common unit, corresponding to the standard deviation of the cues. Stimulus centers in the  $A$  and  $V$  conditions are points on the  $A$  and  $V$  axes corresponding to the means of the Gaussian densities. Stimulus centers in the  $AV$  condition are points (at the centers of the circles indicated) in the  $A$ - $V$  plane whose  $A$  and  $V$  coordinates are equal to the stimulus centers on the  $A$  and  $V$  axes. Thus distances between points in the plane are the Pythagorean sums of the  $A$  and  $V$  distances for the same stimulus pair. In all conditions the distance between a pair of points is the subject's sensitivity for distinguishing between the corresponding consonants.

$$P_{AV}(R_j | S_i) = \sum_{m,n \in \Theta_j} P_A(R_m | S_i) \times P_V(R_n | S_i) \quad (3)$$

An example that illustrates how the model of Post-Labeling Integration can be used to predict scores for the  $AV$  condition is given in Tables 1 through 3. As can be seen, for the hypothesized matrices the improvement in score to be expected from Post-Labeling Integration is: from 62.0% in the auditory condition and 55.3% in the visual condition, to 64.8% in the audiovisual condition. Such small improvements are to be expected when the patterns of confusions in the auditory and visual conditions are highly correlated.

Table 1: *Hypothetical Auditory and Visual Confusion Matrices.* For each matrix, the cell entry gives the proportion of presentations of the stimulus corresponding to the row containing the cell (*A*, *B*, *C*) on which the response corresponding to the column containing the cell (*A*, *B*, *C*) is elicited.

	Auditory Matrix			Visual Matrix		
Stimulus	Auditory Response			Visual Response		
	A	B	C	A	B	C
<i>A</i>	0.60	0.25	0.15	0.42	0.31	0.27
<i>B</i>	0.12	0.50	0.38	0.13	0.53	0.34
<i>C</i>	0.07	0.17	0.76	0.08	0.21	0.71

Table 2: *Predicted Probabilities of Auditory-Visual Label-Pairs.* Each numeric cell entry gives the proportion of presentations of the stimulus corresponding to the row containing the cell (*A*, *B*, *C*) on which the AV label pair corresponding to the column containing the cell (*AA*, *AB*, ..., *CC*) is assumed to be elicited. For example, the label pair *CA* is assumed to be elicited on the proportion 0.049 (= 0.38 x 0.13) of the presentations of stimulus *B*, because on each presentation of this stimulus the auditory response *A* is elicited with probability 0.38 and the visual response *C* is elicited with probability 0.13. Cells in the lowest row identify the stimulus most likely to elicit each label pair. For example, the stimulus most likely to have elicited label pair *AC* is *A* because it elicits the pair on the proportion 0.162 of presentations, while stimuli *B* and *C* elicit the pair only on 0.041 and 0.050 of their presentations.

AV Label Pair Matrix									
Stimulus	AV Label Pair								
	AA	AB	AC	BA	BB	BC	CA	CB	CC
<i>A</i>	0.252	0.186	0.162	0.105	0.078	0.068	0.063	0.046	0.040
<i>B</i>	0.016	0.064	0.041	0.065	0.265	0.170	0.049	0.201	0.129
<i>C</i>	0.006	0.015	0.050	0.014	0.036	0.121	0.061	0.160	0.540
ML Ident.	<i>A</i>	<i>A</i>	<i>A</i>	<i>A</i>	<i>B</i>	<i>B</i>	<i>A</i>	<i>B</i>	<i>C</i>

### Complementary and Redundant Cues

There are two extreme cases of integration that merit further consideration. At one extreme, the sources provide cues that complement one another: each set of cues is inadequate to resolve all stimuli separately, but the combined set permits all stimuli to be distinguished. At the other extreme, the cues provided by the sources are completely redundant with one another: while each set permits stimuli to be distinguished to a certain extent, resolution is enhanced when both sets are used.

In terms of the Pre-Labeling Model, one would expect distances between stimulus centers to be positively correlated across sources in the redundant case (since the salience of cues provided by the two sources is similar for each distinction) but negatively correlated across sources in the

Table 3: *Predicted Audiovisual Confusion Matrix Corresponding to the Maximum Likelihood Mapping.* Each numeric cell gives the predicted proportion of AV presentations of the stimulus corresponding to the row containing the cell (*A*, *B*, *C*) in which the response corresponding to the column containing the cell (*A*, *B*, *C*) is elicited. For example the response *B* is the maximum likelihood assignment for AV label-pairs *BB*, *BC*, and *CB*. When stimulus *C* is presented, these label pairs are assumed to be elicited with probabilities 0.036, 0.121, and 0.160 respectively. Thus response *B* is predicted to be elicited on the proportion 0.316 ( $= 0.036 + 0.121 + 0.160$ ) of presentations of stimulus *C*.

Predicted AV Matrix			
Stimulus	Response		
	A	B	C
<i>A</i>	0.768	0.191	0.041
<i>B</i>	0.234	0.636	0.129
<i>C</i>	0.144	0.316	0.540

complementary case (since small distances for one cue set correspond to large distances for the other, and *vice versa*). An illustrative example of the relative accuracy that can be achieved through the use of complementary and redundant cues is provided in Fig. 4. In this case, in order to achieve comparable accuracy the spacing of stimulus centers for redundant cues must be roughly double that for complementary cues.

Similar conclusions can be reached using the Post-Labeling Model. The case of complementary cues is illustrated by the confusion matrices shown in Table 4. Using only cues provided by Channel 1, it is not possible to distinguish stimuli *A* and *B* very well, but either member of the pair can be distinguished from stimulus *C*; using only cues provided by Channel 2, it is not possible to distinguish stimulus *B* from *C* very well, but either can be distinguished from stimulus *A*. Because the conditions are complementary, the model predicts that in the Combined condition all three stimuli can be distinguished from one another and the predicted score in the AV condition (72.5%) is substantially higher than in the A and V conditions (56.7%). The case of redundant cues is illustrated by the confusion matrices shown in Table 5. As can be seen, the improvements expected from Post-Labeling Integration in this case are substantially smaller than in the complementary-cue case (from 56.7% in the A and V conditions to 58.3% in the AV condition).<sup>1</sup>

<sup>1</sup>The unexpected asymmetry of the predicted confusion matrix in the Combined condition results from arbitrariness in the assignment of label pairs when  $P^*_{AV}(A_m, V_n | S_j) = P^*_{AV}(A_m, V_n | S_i)$ . The probability of a correct response is unaltered if this label pair is mapped into the response corresponding to  $S_i$  or  $S_j$ . The illustrative computation assigned such cases to the stimulus with the alphabetically-earlier designation.



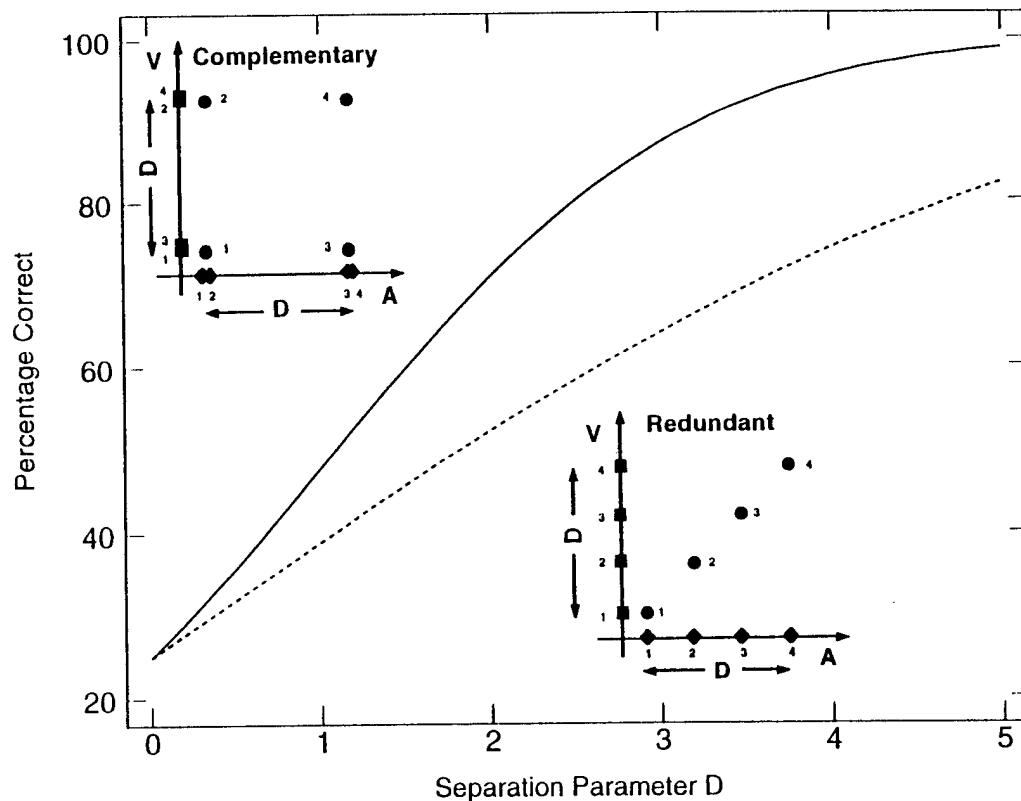


Figure 4: Predictions of the Pre-Labeling Model for accuracy in identifying four stimuli (1, 2, 3, 4) as a function of the perceptual span of unimodal cues. Confusions in the A and V conditions are each assumed to be adequately described by scalar cues (*A* and *V*). Stimulus centers for the *A* (diamonds) and *V* (squares) scalar variables are shown on axes scaled to have a common unit, corresponding to the standard deviation of the cues. Stimulus centers in the AV condition are points in the *A-V* plane marked by the circles. In the redundant-cue case (dashed curve) stimulus centers are assumed to be uniformly spaced over the *A* and *V* spans. In the complementary-cue case (solid curve), pairs of stimulus centers are grouped at the ends of the span, with different groupings for the *A* and *V* cues.

### Applications to Intelligibility Prediction

Braida (1991) has applied these models to data obtained in five modern studies of audiovisual consonant identification. The predictions of the Post-Labeling Model, although accurate in some cases, systematically underestimate the multimodal accuracy in all cases. By contrast, the predictions of the Pre-Labeling Model are both higher than those predicted by the Post-Labeling Model and closer to observed values. On average, predicted audiovisual scores were only 0.5 percentage point less than observed scores, and cannot be distinguished from them statistically. Moreover, the patterns of residual confusions seen for audiovisual presentation conditions are in reasonably good agreement with the predictions of the Pre-Labeling Model.

The same analysis can also be applied to the integration of cues from different channels in the same sensory system, e.g., from low- and high-frequency bands of filtered speech. The amount of

Table 4: *Hypothetical Complementary Confusion Matrices*

	Channel 1			Channel 2			Combined		
Stimulus	Response			Response			Response		
	A	B	C	A	B	C	A	B	C
<i>A</i>	0.75	0.20	0.05	0.70	0.20	0.10	0.665	0.285	0.050
<i>B</i>	0.65	0.25	0.10	0.10	0.25	0.65	0.090	0.810	0.100
<i>C</i>	0.10	0.20	0.70	0.10	0.15	0.75	0.030	0.270	0.700

Table 5: *Hypothetical Redundant Confusion Matrices*

	Channel 1			Channel 2			Combined		
Stimulus	Response			Response			Response		
	A	B	C	A	B	C	A	B	C
<i>A</i>	0.60	0.25	0.15	0.60	0.25	0.15	0.840	0.062	0.097
<i>B</i>	0.25	0.50	0.25	0.25	0.50	0.25	0.438	0.250	0.313
<i>C</i>	0.15	0.25	0.60	0.15	0.25	0.60	0.278	0.062	0.660

relevant data available on such integration is small (e.g., Miller and Nicely, 1955) but the results seem consistent with the finding for audiovisual consonant identification. Miller and Nicely tested two pairs of filtering conditions that, in combination, correspond roughly to their wideband (0.2-5.0 kHz) condition. One pair consisted of a low-frequency band from 0.2-2.5 kHz and a high-frequency band from 2.5-5.0 kHz; the other consisted of the slightly overlapping 0.2-1.2 and 1.0-5.0 kHz bands. As can be seen in Table 6, for the 0.2-2.5/2.5-5.0 case the predictions of the Post-Labeling Model are lower than those of the Pre-Labeling Model and also lower than the scores observed in the wide-band condition. The predictions of the Pre-Labeling Model exceed observed scores only very slightly. In the 0.2-1.2/1.0-5.0 case, the prediction of the Post-Labeling Model is only slightly lower than the observed score while that of the Pre-Labeling Model exceeds the observed score by 6.0 percentage points. These results are consistent with those seen for the 0.2-2.5/2.5-5.0 case if the roughly one-third octave band overlap is taken into account.

Table 6: *Consonant Identification Scores Reported by Miller and Nicely (1955) for Filtered Bands of Speech and Scores Predicted by the Integration Models*

Bands (kHz)	Percentage Correct				
	Lowpass	Highpass	Wideband	Pre-Label.	Post-Label.
0.2-1.2 1.0-5.0	57.2	73.1	83.3	89.3	83.2
0.2-2.5 2.5-5.0	72.8	38.1	83.3	83.5	77.9

These models can also be extended to predict the effect of changing signal-to-noise ratios on intelligibility. When signals are observed twice in statistically independent noise, the effect of the noise can be reduced if the observations are combined appropriately. Under optimum conditions, the effective signal-to-noise ratio is improved by  $\sqrt{2}$  for each doubling of the number of observations. Data on the effect of changing the signal-to-noise ratio on consonant reception were obtained by Miller and Nicely (1955) who used white noise with S/N from -18 to +12 dB (in 6 dB steps). In the Pre-Labeling Model, the effect of multiple observations of cues in the presence of independent noise is the same as increasing the distance between stimulus centers: all distances increasing by  $\sqrt{2}$  for each doubling of observations or increasing by 2 for each increase of S/N by 6 dB. The predictions of the Pre-Labeling Model are compared to observed scores in Fig. 5. Post-Labeling Model predictions (Fig. 6) were made by assuming that responses at a given S/N were based on four statistically independent observations made at a 6 dB lower S/N. As can be seen, both models make reasonable predictions for the effects of changing S/N on consonant identification, at least over part of the range of S/N tested by Miller and Nicely. The Pre-Labeling Model tends to overestimate scores at high S/N, but makes fairly good predictions at low S/N. The Post-Labeling Model is somewhat more accurate at high S/N, but less capable of accounting for S/N changes at low S/N.

According to the assumptions of the Pre-Labeling Model, if the Miller and Nicely (1955) confusion matrices are represented using conventional multidimensional scaling techniques, the shape of the configuration of points representing the consonants should be relatively invariant to changes in S/N. Such an invariance was observed by Shepherd (1972) using clustering techniques. He found that S/N determined the overall level of confusion but had little effect on the pattern of confusion (at least over the range -12 to +6 dB). In spatial solutions, the relative distance between consonants was invariant to changes in S/N. This invariance was found to be much more evident in the estimated distances than in the raw confusion matrices. Although highly suggestive, Shepherd's analysis did not indicate how the scale of the configuration varied depending on S/N. Interestingly, Shepherd also reported some evidence for the relation between the low-pass and high-pass cue spaces and the broadband cue space. In the high-pass filtering conditions, consonants separated horizontally in the configuration were most confused, while in the most extreme low-pass condition the prevalent confusions corresponded to vertical separations. Based on this observation, Shepherd conjectured that the vertical and horizontal dimensions of the configuration he derived reflect discriminations among high frequencies and low frequencies, respectively.

### **Relation to Articulation Theory**

The use of the Pre-Labeling Model to relate identification scores derived from different presentation conditions is similar to the use of Articulation Theory to predict the effect of filtering on intelligibility (e.g., ANSI, 1969). There are, however, substantial differences. Since the predictions of Articulation Theory are reasonably accurate within its domain of applicability (e.g., accounting for the effects of presentation level, linear filtering, and additive noise), it is of interest to contrast the Articulation Theory approach with that of the integration models.

In Articulation Theory, the intelligibility of a processed speech signal reflects the accumulation of an abstract quantity measured by the Articulation Index. The contribution to the Index by each spectral band of speech is determined by the proportion of speech band levels that are audible (up to a

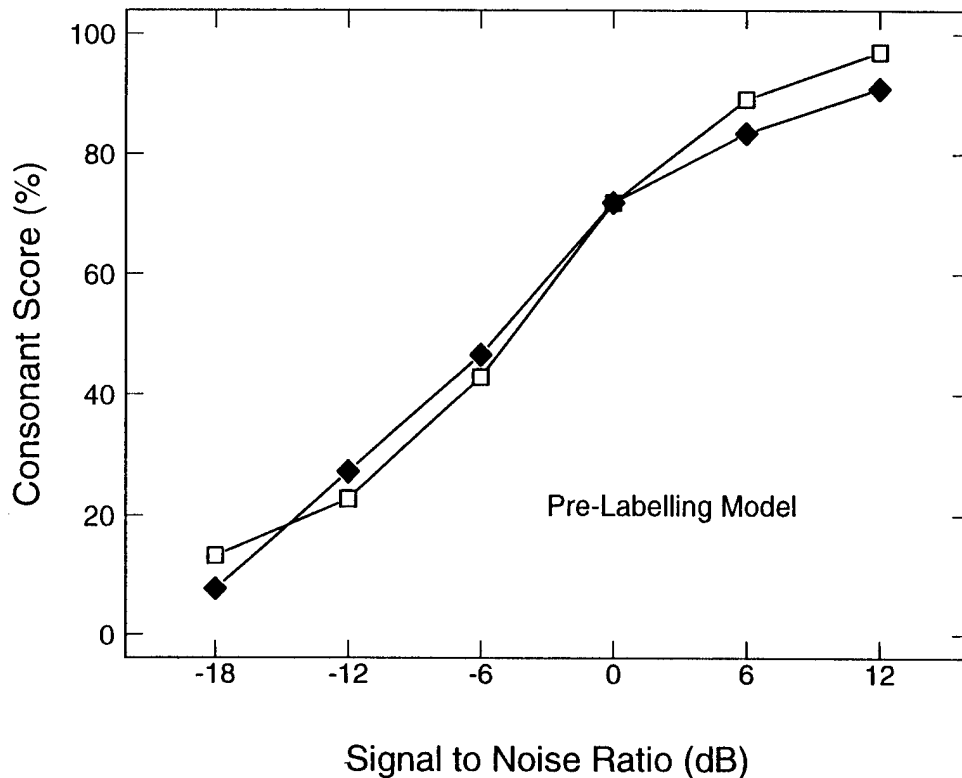


Figure 5: Observed (filled diamonds) and predicted (unfilled squares) identification scores for 16 consonants presented in broadband noise. Observed scores were reported by Miller and Nicely (1955). Predicted scores were computed according to the assumptions of the Pre-Labeling Model. A three-dimensional description of the consonant confusion matrix obtained in the 0 dB S/N condition was determined using the iterative procedures described by Braida (1988). For the other S/N conditions all distances were increased by a factor of 2.0 for each 6 dB increase in S/N and reduced by a factor of 2.0 for each 6 dB decrease in S/N. The relative locations of stimulus and response centers were not changed as S/N varied. Predicted identification scores were computed by integrating the densities centered at each stimulus center within response regions determined by the locations of the response centers.

maximum of 30 dB) and the relative importance of the band for communication. When two non-overlapping bands are summed, the Index for the combination is the sum of the indices for the individual bands. In Articulation Theory, the effect of varying S/N is to change the proportion of speech band levels that are audible in each band. The relationship between the cumulative Articulation Index and the intelligibility score in a given speech test is determined by a nonlinear monotonic function dependent on the difficulty of the intelligibility test and the redundancy of the test materials. When speechreading is available, Articulation Theory assumes that the relation between Articulation Index for the acoustic signal ( $AI_A$ ) and that for the audiovisual signal ( $AI_{AV}$ ) is given by a monotonic function that is independent of the acoustic signal; all signals characterized by the same value of  $AI_A$  are predicted to yield the same value of  $AI_{AV}$ .

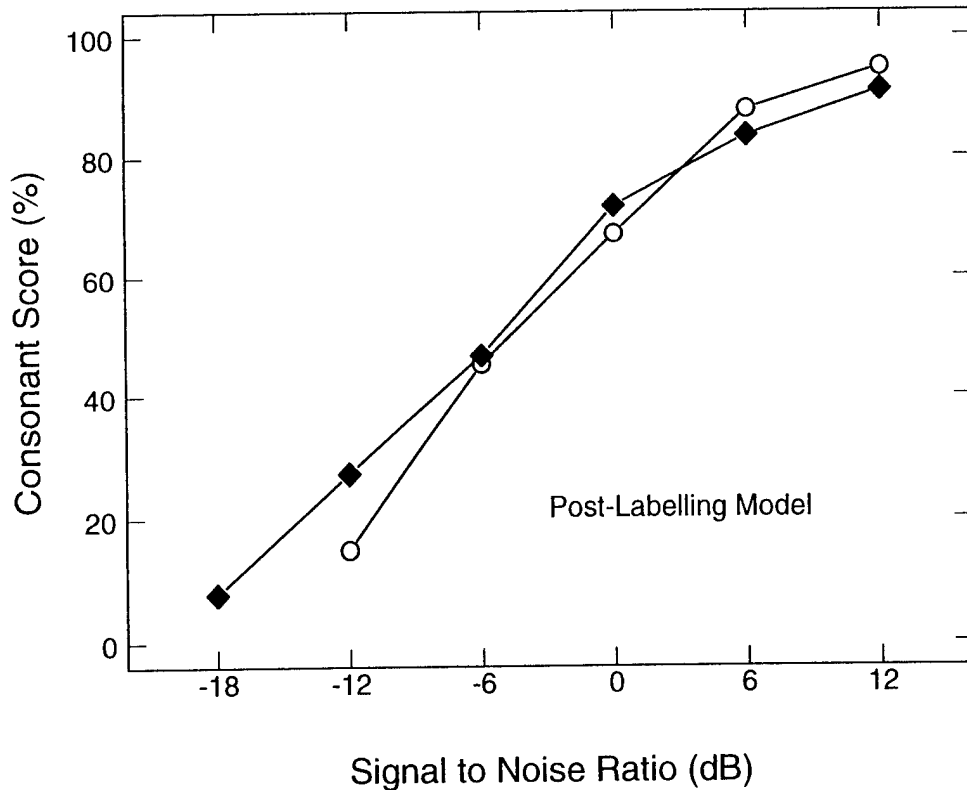


Figure 6: Observed (filled diamonds) and predicted (unfilled circles) identification scores for 16 consonants presented in broadband noise. Observed scores were reported by Miller and Nicely (1955). Predicted scores were computed according to the assumptions of the Post-Labeling Model. At each S/N the observed confusion matrix was used to predict the identification score at the next higher S/N. Responses at the higher S/N were assumed to be based on four statistically independent observations made at the lower S/N.

The account of the dependence of the intelligibility of consonants subject to the manipulations considered in this paper that is provided by Articulation Theory differs from that provided by the integration models. Consider first the effect of filtering. According to Articulation Theory, the relation between the intelligibility expected when two bands are combined depends only on the AI values, and hence the intelligibility scores, for the individual bands: any pair of non-overlapping bands with the same pair of AI values (or individual band scores) are predicted to yield the same score when the bands are summed.<sup>2</sup> The integration models predict that the combined score should depend on the detailed perceptual confusions that characterize each of the bands.<sup>3</sup> In general, for a given pair of band intelligibility scores, bands that are characterized by similar confusion patterns (or provide redundant

<sup>2</sup>To predict the intelligibility of the combined band, one must also know the function that specifies the dependence of intelligibility on the Articulation Index for the materials. In general, this requires knowing the intelligibility for conditions with higher values of the Articulation Index than for the individual bands, e.g., for bands broader than either of the bands that are combined.

<sup>3</sup>This prediction does not require knowledge of the intelligibility for any other bands.

cues) are expected to yield lower scores when combined than bands characterized by complementary confusion patterns (in which, for example, stimulus *A* is confused with *B* but not with *C* in one band, and *A* is confused with *C* but not with *B* in the other band). In this regard, it is interesting to note that although Articulation Theory generally makes good predictions when adjacent bands are combined, it often fails when the bands are well separated (e.g., Kryter, 1962).

Articulation Theory accounts for improvement of intelligibility scores when S/N is increased (or presentation level is increased in the absence of noise) in terms of the increase in the range of speech band levels that is audible to the listener. By contrast, integration models account for the improvement by equating the effect of increases in S/N to multiple observations in uncorrelated noise. Both accounts must deal with the fact that for a filtered band of speech presented in a noise background, increases in S/N beyond a certain point fail to yield significant intelligibility increases, even though scores are far from perfect. In Articulation Theory, this is modelled by assuming that the contribution of any band to intelligibility is maximal when the highest 30 dB range of band levels is audible. Although not discussed above, integration models could account for this effect by postulating the existence of an additive internal noise whose level was fixed, independent of the external S/N. As S/N increased, intelligibility would ultimately be limited by this internal noise rather than external noise.

According to Articulation Theory, the availability of speechreading improves the intelligibility that can be achieved with the acoustic signal because the  $AI_{AV}$  is increased relative to  $AI_A$  by an amount that depends on  $AI_A$  (being largest for small values of  $AI_A$ ) but is independent of other properties of the acoustic signal. Two acoustic signals, e.g. low-pass and high-pass filtered speech, that are equally intelligible without speechreading are thus predicted to be equally intelligible with speechreading. According to the integration models, the improvement in intelligibility associated with the availability of speechreading should, as in the case of filtered bands of speech, depend on the specific confusion patterns for the acoustic and visual signals. The relatively large improvements seen when speechreading supplements a severely low-passed or noise-corrupted acoustic speech signal result from the complementary nature of the visual and acoustic cues in these cases; visual cues permit accurate identification of place of articulation but not of consonant voicing, while acoustic cues permit accurate identification of voicing but not of place. For less degraded acoustic signals, the improvement is predicted to be less because, since place distinctions can be partially made on the basis of the acoustic signal, the visual signal provides redundant rather than complementary cues.

Although Articulation Theory can be used to predict the intelligibility of a wide variety of materials (e.g., syllables, words, sentences), the integration models make predictions only for small, closed sets of speech segments. To extend these models to a wider range of speech materials, models for integration appropriate for vowels, and models that relate syllable identification to consonant and vowel identification (e.g., Boothroyd, 1988; Rabinowitz et al., 1992) would also be required. On the other hand, by basing intelligibility predictions on integration models it should be possible to obtain new insights into the effectiveness of speechreading supplements whose acoustic intelligibility is practically negligible, such as those derived from fundamental frequency or the amplitude envelopes of filtered bands of speech. Although such signals provide little intelligibility for sentences or large sets of words, they can be used to identify consonants at levels well above chance. Unfortunately the necessary triads of confusion matrices for these signals are not available.

## Conclusion

Additional research is needed to provide more incisive tests of the integration models described in this paper. The description of unimodal sensitivity used in the Pre-Labeling Model needs to be evaluated in more detail, on data from individual subjects. These evaluations should test the notion that model sensitivity is independent of those changes in response center locations that can be effected by changes in instructions, presentation probabilities and payoffs, as well as determine the extent to which the model provides a statistically adequate description of confusion matrices. This will entail the development of more robust techniques for estimating model parameters.

In a similar vein, the integration models need to be tested over a larger range of stimuli in both the auditory-visual and auditory-auditory cases. Even the relatively plentiful studies of audiovisual integration have been almost entirely restricted to the identification of segments distinguished by consonants, and these generally employ a single vowel context. To evaluate the integration models more adequately, data on audiovisual reception of vowels and of consonants in a wider range of vowel contexts is required. Similarly, in the auditory-auditory case, data on segment reception under a variety of bandpass filtering conditions is needed to extend the tests based on the data of Miller and Nicely (1955).

In addition to further tests of the integration models, it is also appropriate to consider further applications of the models to understanding speech perception. For example, in the field of physiological modelling, recent studies of the responses of populations of neural elements in the auditory system to speech stimuli have spurred attempts to understand the relation between neural responses and speech reception. These modelling efforts generally have not considered limitations on the ability to make cross-frequency comparisons of auditory stimuli such as those seen in studies of spectral-shape discrimination and profile analysis (e.g., Durlach et al., 1986, Farrar et al., 1987; Green, 1988). The effects of these limitations can be studied by modelling the physiological interpretation of frequency-specific ranges of neural elements and using the integration models to combine interpretations across neural ranges.

Finally, many users of hearing aids, cochlear implants, and tactile aids often derive substantial communication benefit only when the prosthesis is used together with speechreading. The design of such prostheses is thus a matter of specifying the input to one modality when the performance of the device is ultimately evaluated on both unimodal and multimodal performance. Optimizing prosthesis design is likely to proceed most efficiently if intermodal integration is explicitly taken into account. To the extent that this integration can be modelled, it should be possible to apply these analytical techniques in reverse fashion: given a target level of performance for the audiovisual condition and a specified confusion matrix for the visual condition, determine an auditory confusion matrix consistent with the target level of accuracy. The prosthesis design problem can then be confined to achieving a specified auditory confusion matrix.

## Acknowledgement

This work was supported by the N. I. H. (Grants R01-DC00117 and P01-DC00361).

## References

- ANSI. *American National Standard Methods for the Calculation of the Articulation Index*, ANSI S3.5-1969. New York: American National Standards Institute (1969).
- Boothroyd, A., and Nitttrouer, S. "Mathematical treatment of context effects in phoneme and word recognition," *J. Acoust. Soc. Am.*, 84, 101-114 (1988).
- Braida, L.D. "Development of a model for multidimensional identification experiments," *J. Acoust. Soc. Am.*, 84, S142 (1988).
- Braida, L.D. "Crossmodal integration in the identification of consonant segments," *Quart. J. Exper. Psych.*, 43A, 647-677 (1991).
- Breeuwer, M., and Plomp, R. "Speechreading supplemented with formant frequency information from voiced speech," *J. Acoust. Soc. Am.*, 77, 314-317 (1985).
- Breeuwer, M., and Plomp, R. "Speechreading supplemented with auditorily presented speech parameters," *J. Acoust. Soc. Am.*, 79, 481-499 (1986).
- Durlach, N.I., Braida, L.D., and Ito, Y. "Towards a Model for the Discrimination of Broadband Stimuli," *J. Acoust. Soc. Am.*, 80, 63-72 (1986).
- Farrar, C.L., Reed, C.M., Durlach, N.I., Delhorne, L.A., Zurek, P.M., Ito, Y., and Braida, L.D. "Spectral-shape Discrimination. I. Results from Normal-hearing Listeners for Stationary Broadband Noise," *J. Acoust. Soc. Am.*, 81, 1085-1092 (1987).
- Grant, K.W., and Braida, L.D. "Evaluating the Articulation Index for Audiovisual Input," *J. Acoust. Soc. Am.*, 89, 2952-2960 (1991).
- Grant, K.W., Braida, L.D., and Renn, R.J. "Single-band envelope cues as an aid to speechreading," *Quart. J. Exper. Psych.*, 43A, 621-645 (1991).
- Grant, K.W., Braida, L.D., and Renn, R.J. "Auditory supplements to speechreading: combining amplitude envelope cues from different spectral regions of speech," submitted to *J. Acoust. Soc. Am.* (1993).
- Green, D.M. Profile Analysis: Auditory Intensity Discrimination. New York: Oxford University Press (1988).
- Kryter, K.D. "Validation of the Articulation Index," *J. Acoust. Soc. Am.*, 34, 1698-1702 (1962).
- Miller, G.A., and Nicely, P. "An analysis of perceptual confusions among some English consonants," *J. Acoust. Soc. Am.*, 27, 328-352 (1955).



Rabinowitz, W.M., Eddington, D.K., Delhorne, L.A., and Cuneo, P.A. "Relations among different measures of speech reception in subjects using a cochlear implant," *J. Acoust. Soc. Am.*, 92, 1869-1881 (1992).

Shepherd, R.N. "Psychological representation of speech sounds," in E.E. David and P.B. Denes (Eds.) Human communication: A unified view. New York: McGraw-Hill (1972).

**Appendix A. Program of Symposium**

**National Academy of Sciences/National Research Council**

**COMMITTEE ON HEARING,  
BIOACOUSTICS, AND BIOMECHANICS**

**Symposium on  
SPEECH COMMUNICATION METRICS  
AND  
HUMAN PERFORMANCE**

**JUNE 3-4, 1993**

**National Academy of Sciences Auditorium  
21st and C Streets N.W.  
Washington DC**

## **PROGRAM**

**THURSDAY, JUNE 3, 1993**

**8:30            REGISTRATION**

**8:45            WELCOME**

Henning von Gierke, Chair of CHABA  
Committee on Hearing, Bioacoustics, and Biomechanics

**SESSION CHAIR: NEAL VIEMEISTER**

**9:00            DEVELOPMENT OF SPEECH INTELLIGIBILITY  
MEASURES AND THE ANSI STANDARD**

Mones Hawley  
Jack Fawcett Associates  
Bethesda, MD

**9:45            PROPOSED REVISION OF ANSI STANDARD FOR  
DETERMINING THE ARTICULATION INDEX**

Patrick Zurek, Massachusetts Institute of Technology  
Chaslav Pavlovic, University of Iowa

**10:30          COFFEE BREAK**

**11:00          SOURCES OF VARIABILITY AFFECTING SPEECH  
INTELLIGIBILITY TESTS**

David Pisoni  
Indiana University

**11:45          \*\*\*COMMUNICABILITY MEASURES OF NARROW-BAND  
DIGITAL VOICE COMMUNICATION SYSTEMS**

John Terdelli and Elizabeth Kreamer  
ARCON Corporation  
Waltham, MA

**\*\*\*Manuscript was not provided to CHABA.**

**THURSDAY, JUNE 3, 1993**

**SESSION CHAIR: THOMAS J. MOORE**

- 2:00        SPEECH INTELLIGIBILITY EFFECTS IN A DUAL  
TASK ENVIRONMENT**  
David Payne  
SUNY, Binghamton
- 2:45        THE EFFECTS OF SPEECH INTELLIGIBILITY ON  
MILITARY PERFORMANCE**  
Georges Garinther, Leslie Whitaker, and Leslie Peters  
Army Research Laboratory  
Aberdeen, MD
- 3:00        COFFEE BREAK**
- 3:45        THE EFFECTS OF MESSAGE COMPLEXITY ON  
PERFORMANCE**  
Andrew Rose  
American Institutes for Research  
Washington, DC
- 4:30        A VOICE COMMUNICATION EFFECTIVENESS TEST**  
Richard McKinley  
Wright-Patterson Air Force Base, OH
- 5:15        ADJOURN**
- 5:30        RECEPTION**

**FRIDAY, JUNE 4, 1993**

**SESSION CHAIR: JUDY DUBNO**

- 8:30            INDIVIDUAL DIFFERENCES IN SPEECH PERCEPTION BY  
EYE AND EAR**  
Charles S. Watson  
Indiana University
- 9:15            SEQUENCE COMPARISON TECHNIQUES CAN BE USED TO  
STUDY SPEECH PERCEPTION**  
Lynn Bernstein  
Gallaudet University
- 10:00           COFFEE BREAK**
- 10:15           APPLICATIONS OF GENERALIZABILITY THEORY TO  
MEASUREMENT OF INDIVIDUAL DIFFERENCES IN  
SPEECH PERCEPTION**  
Marilyn Demorest  
University of Maryland, Baltimore
- 11:00           MODELING AUDITORY AND VISUAL CONTRIBUTION TO  
SPEECH INTELLIGIBILITY**  
Louis Braida  
Massachusetts Institute of Technology
- 11:45           \*\*\*DIGITAL SPEECH PROCESSING: INTERACTION OF  
SCIENCE, TECHNOLOGY, AND THE MARKETPLACE**  
Stephen Levison and Lawrence Rabiner  
AT&T Bell Laboratories  
Murray Hill, NJ
- 12:30           DISCUSSION**
- 1:00            ADJOURN**

**\*\*\*Manuscript was not provided to CHABA.**